Big Data Privacy Challenges and Techniques

Nitin Kumar Agrawal
Department of Computer Applications,
Faculty of Engineering & Science,
Mangalayatan University

Aprna Tripathi, PhD
Department of Computer Applications,
Faculty of Engineering & Science,
Mangalayatan University

ABSTRACT

Big Data deals with a diverse and vary large set of data that includes images, video, audio, structured, unstructured as well as semistructured data. Big Data is getting bigger day by day because of voluminous amount of data that is being collected, processed and stored. It has created new challenges to preserve privacy and hence we have many privacy challenges and privacy techniques to deal with in Big Data era. There is always a potential risk to privacy in Big Data and as data is everywhere, privacy risk moves with data all the way. There are privacy requirements in Big Data and these needs to be handled very carefully otherwise risk associated with it will definitely increase.

Keywords

Big Data, structured, unstructured and semi structured data, privacy challenges, privacy techniques.

1. INTRODUCTION

The Big Data becomes challenge for traditional systems not merely because of its size. Size is a challenging point, but challenge could also arise because of speed at which the Big Data is coming and also because it is unstructured, it could contain data items of various formats. Security aims to reduce risk. To reduce risk, specific information about a resource is often called for. As soon as the resource is related to entities such as individuals and corporates, privacy concerns creep in [6].

2. WHAT IS BIG DATA?

'Big Data' refers to novel ways in which organizations, including government and businesses, combine diverse digital datasets and then use statistics and other data mining techniques to extract from them both hidden information and surprising correlations. While Big Data promises significant economic and social benefits, it also raises serious privacy concerns [1]. The term Big Data is very misleading sometimes as gives the impression that after certain size the data is big and below a certain size the data is small.

The below table gives the information to measure the size of the data:

Table 1: Size of Data

Name	Symbol	Value
Kilobyte	KB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}
Zettabyte	ZB	10^{21}
Yottabyte	YB	10^{24}

The question here is from which point the Big Data starts? Though the answer is not that simple, however, the answer is, it depends [9]. The Big Data could start from any point. There is no definitive definition for Big Data, however it is mostly defined as follows: "Big Data is a data that becomes difficult to be processed because of its size using the traditional systems."

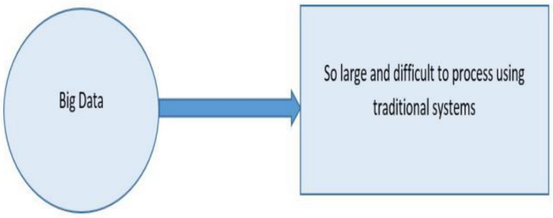

Figure 1: What is big Data?

When we say that Big Data is so large and difficult to process by using traditional systems, what exactly we mean? Consider the following examples:

Let's say that you have created a document of 100MB and you want to share it with your colleague and you are unable to send it by email. So, this becomes Big Data for you because you are unable to use traditional methods with this document because of its large size.

Let's say you have an image file of 100GB and you are unable to display it on your computer screen in real time because of the size of this image file. So, this becomes Big Data for you in this context.

Let's say you have a video file of 100TB and you are unable to edit it using your editing software. So, this video file becomes Big Data for you.

Therefore, it completely depends on the capabilities of the system to process the data. So, the term Big Data is related to the capabilities of the system. At a higher level the term is related to the capabilities of the organization. For example: consider two companies, Company A and Company B. Let's say a stream of data is coming into these two companies. Assume that the data set consists of different unstructured data like text, audio, video etc., and 500TB of data is coming on daily basis. This data set could be a Big Data for company A and not a Big Data for company B. Because it depends on the capabilities of these two companies. [Here we assume that company B is all set to digest this volume and variety of data while at this velocity and variety company A is not prepared yet.

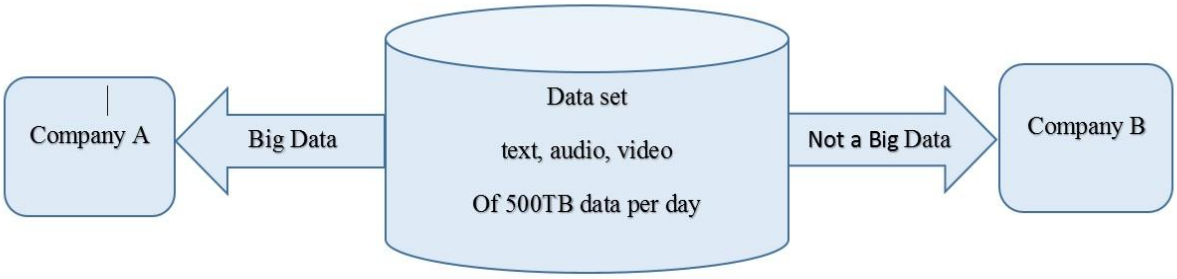

Figure 2: When a data set is Big Data and when not.

Big data is a term that describes the large volume of data – both structured and unstructured. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

3. TYPES OF DATA

Data exists in various types [10], consider Figure 3, and they can be defined as follow:

3.1 Structured data

Structured data refers to information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations. Structured data is information, usually text files, displayed in titled columns and rows which can easily be ordered and processed by data mining tools. This could be visualized as a perfectly organized filing cabinet where everything is identified, labeled and easy to access.

3.2 Unstructured data

Whereas unstructured data is essentially the opposite of the structured data. The lack of structure makes compilation a time and energy-consuming task.

3.3 Semi-structured data

Semi-structured data lies somewhere between the two. It is not organized in a complex manner that makes sophisticated access and analysis possible; however, it may have information associated with it, such as metadata tagging that allows elements contained to be addressed.

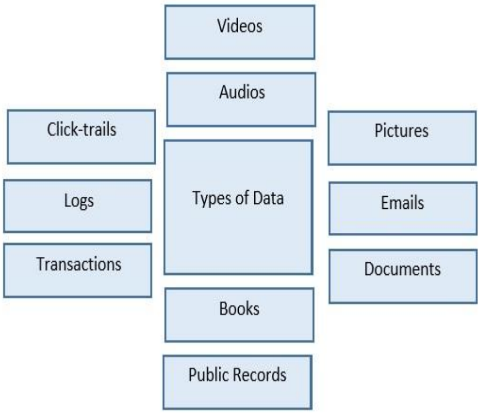

Figure 3: Types of data.

3.4 Example

A Word document is generally considered to be unstructured data. However, you can add metadata tags in the form of keywords and other metadata that represent the document content and make it easier for that document to be found when people search for those terms -- the data is now semi-structured. Nevertheless, the document still lacks the complex organization of the database, so falls short of being fully structured data.

4. BIG DATA 'V' ATTRIBUTES: CHARACTERISTICS OF BIG DATA

The famous 3 'V's of Big Data that are volume, variety and velocity has now been increased to few or more V's that includes variability, complexity and others. The 'V' attributes are not limited to 5V's only. Here we are considering the above V attributes of Big Data.

4.1 Volume

Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden. Volume is of course a problem. The data keeps on getting accumulated and the file becomes too large to be handled by a traditional system. Example: Facebook is generating 25 TB of data daily, just imagine the size of the files that are there at the beginning of time.

4.2 Velocity

Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. The velocity refers to the speed at which the data is coming. Example: The scientific experiments that scientists do at the atomic reactors where they do the collision of the subatomic particles, around 40TB data is coming in within 1second, that is a very high speed.

4.3 Variety

Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. In traditional systems the data is structured and it is stored in well planned tables. Each table has specific columns and each column could accept values of specific data types. However, in case of Big Data, this 'V' that is 'variety' creates problems sometimes. It may include data items of various formats that could have audio files, video files, unstructured data etc. It becomes challenging for a traditional system to handle this type of data.

4.4 Variability

In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data.

4.5 Complexity

Today's data that we have and we work with comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

5. IMPORTANCE OF BIG DATA

The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable the following:

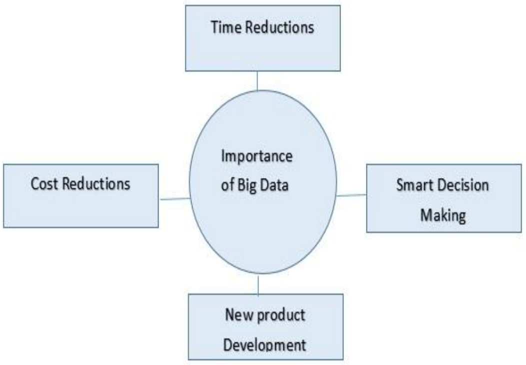

Figure 4: **Importance of Big Data**

6. BIG DATA GENERATION POINTS

Big data is generated through various sources. The following image shows various points from where data is coming at a very high speed resulting into a pool of Big Data with large files. Sometimes the data is not even structured.

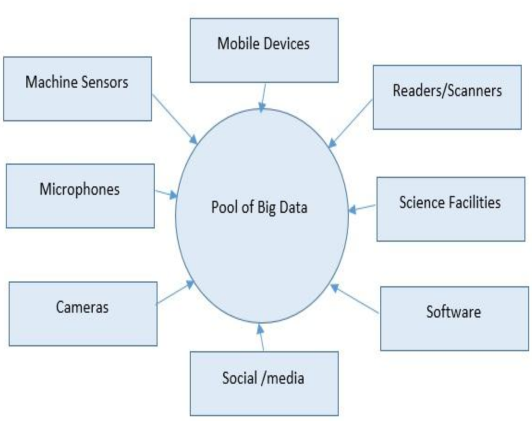

Figure 5: Big Data Generation Points

7. WHAT IS PRIVACY?

Privacy is an essential way we seek to protect ourselves and society against arbitrary and unjustified use of power, by reducing what can be known about us and done to us, while protecting us from others who may wish to exert control. Privacy is essential to who we are as human beings, and we make decisions about it every single day. It gives us a space to be ourselves without judgment, allows us to think freely without discrimination, and is an important element of giving us control over who knows what about us [2]. Privacy is essential in many ways. Personal elements in life such as intimacy, friendship, role play, and creative experimentation need to remain private. However, IT-enabled and IT-networked world of today is posing challenges in maintaining privacy. In data world, privacy typically refers to Personally Identifiable Information [7].

We can say that privacy is a fundamental right for every human being, essential to autonomy and the protection of human dignity, serving as the foundation upon which many other human rights are built. Privacy enables us to create barriers and manage boundaries to protect ourselves from unwarranted interference in our lives, which allows us to negotiate who we are and how we want to interact with the world around us. Privacy helps us establish boundaries to limit who has access to our bodies, places and things, as well as our communications and our information. Privacy is the right to be let alone, or freedom from interference or intrusion. Information privacy is the right to have some control over how your personal information is collected and used [3].

Privacy can simply be defined as the right to be left alone. A society in which there is a total lack of privacy would be intolerable; but then again a society in which there is a total privacy would be no society at all. Privacy is the right of people to make personal decisions regarding their own intimate matters, it is the right of people to lead their lives in a manner that is reasonably secluded from public scrutiny. In general, the right of people to be free from such things as secretsurveillance and to determine whether, when, how, and to whom, one's personal or organizationalinformation is to be revealed.

7.1 Privacy Categories

Privacy may be divided into four categories

Table 2: Privacy Categories

Privacy Category	**Information**
Physical	Restriction on others to experience a person or situation through one or more of the human senses.
Informational	Restriction on searching for or revealing facts that are unknown or unknowable to others.
Decisional	Restriction on interfering in decisions that are exclusive to an entity.
Dispositional	Restriction on attempts to know an individual's state of mind.

7.2 Why does privacy matter in Big Data?

In today's world where technological innovations are happening at the speed-of-light, information privacy is becoming more complex by the minute as more data is being collected and exchanged. As the technology gets more invasive, so do the uses of data. And that leaves organizations facing an incredibly complex risk matrix for ensuring that personal information is protected. As a result, privacy has fast-emerged as perhaps the most significant protection issue in the global world of information technology. Privacy aims to protect information that is considered personal and should not be shared without informed consent. For example, it is a very common practice to understand drug effectiveness on patients by studying medical records. While such a study is essential to understand and develop more effective treatment, Personal Identifiable Information part of the medical record needs to be removed by anonymization to protect individual patients [7].

7.3 Effect of Big Data on Privacy

Big data has impacted privacy to a very large extent that the countries are changing their privacy laws. Whether it is related to consumers or to the companies.

8. PRIVACY VERSUS SECURITY

Does privacy and security are the same thing? Not really. But they are highly related to each other. Data privacy is focused on the use and governance of personal data, things like putting policies in place to ensure that consumers' personal information is being collected, shared and used in appropriate ways. Security focuses more on protecting data from malicious attacks and the exploitation of stolen data for profit. While security is necessary for protecting data, it's not sufficient for addressing privacy.

8.1 What is information privacy?

Information privacy is the ability of an individual or group to stop information about themselves from becoming known to people other than those they choose to give the information to. Privacy is sometimes related to anonymity although it is often most highly valued by people who are publicly known. Privacy can also be seen as an aspect of security, one in which there are trade-offs between the interests of one group and another can become particularly clear.

8.2 Techniques for Manipulating Personal Information

Privacy of individuals whose data are being collected and analyzed is at risk. Very often, individuals are not even aware their data is used without individuals being aware about it. Here are some of the techniques that governments and other institutions use to manipulate personal information.

8.2.1 Data Merging

This iswhen a databases is merged with another database. **For example**: A database with your driver's license details is merged with a database about your car registration. Or it is when a database with your University subjects is merged with the Department of Immigration.

8.2.2 Data Matching

Thisis when information on a discrete database is used to match similar records on another database. For example: The Taxation office matching banking records. This is similar to data merging as you gave information to one department or Organisation to be used in one context but you didn't give permission for that information to be used in another context.

8.2.3 Data Mining

This is the technique most favored by the Private Sphere. Private companies, such as Microsoft, can use the Information that they gather through systems such as '

Hotmail to uncover social trends that places them in a better position to market goods and services to people (like Google that uses targeted adds). If a company can understand broad commercial and social trends, through data mining large volumes of information, then they have a great competitive advantage over other companies.

9. PRIVACY ISSUES AND CHALLENGES IN BIG DATA

9.1 One important problem in Big Data is ensuring privacy of linked data, e.g. social networks, where people are linked to other people, and relational data, where different types of entities maybe linked to one another. Reasoning about privacy in such data is tricky since information about an individual may be leaked through links to other individuals.

9.2 Another interesting problem is that of releasing sequential releases of the same data over time. Attackers may link individuals across releases and infer additional sensitive information that they could not have from a single release.

9.3 Finally, as the data we deal with become extremely high dimensional, we need to develop techniques that can protect privacy while guaranteeing utility. Understanding theoretical trade-offs between privacy and utility is an important open area for research.

9.4 While big data can yield extremely useful information, it also presents new challenges with respect to how much data to store, how much this will cost, whether the data will be secure, and how long it must be maintained.

9.5 Among the major challenges of big data is preserving individual privacy. As we go about our everyday lives, we leave behind digital footprints that, when combined, could denote unique aspects about we that would otherwise go unnoticed [6].

10. PRIVACY TECHNIQUES

There are so many privacy techniques. We will consider the following privacy techniques:

10.1 Anonymization

Data Anonymization is the process of destroying tracks, or the electronic trail, on the data that would lead an eavesdropper to its origins. An electronic trail is the information that is left behind when someone sends data over a network [8]. For example: forensic experts can follow the data to figure out who sent it. This is often done in criminal cases, but sometimes companies undermine user privacy in order to track user data [6].

10.2 Aggregation

Aggregation reduces disclosure risks by turning atypical records which generally are most at risk into typical records. For example, there may be only one person with a particular combination of demographic characteristics in a city, but many people with those characteristics in a state. Releasing data for this person with geography at the city level might have a high disclosure risk, whereas releasing the data at the state level might not.

10.2.1 Problem with Aggregation

Aggregation is very similar to K-anonymity. Unfortunately, aggregation makes analysis at finer levels difficult and often impossible, and it creates problems of ecological inferences (relationships seen at aggregated levels do not apply at disaggregated levels).

10.3 Suppression

In suppression sensitive values can be deleted from the data. Entire variables or just at-risk data values can be suppressed. Suppression of particular data values generally creates data that are missing because of their actual values, which are difficult to analyze properly. For example, if incomes are deleted from an employee dataset just because they are large, estimates of the income distribution based on this data will be biased.

10.4 Data Swapping

In swapping data values for selected records can be swapped. **For example:** switch values of age, race, and gender for at-risk records with those for other records to discourage users from matching, since matches may be based on incorrect data.

10.4.1 Problem with Swapping

Swapping is used extensively by government agencies. It is generally presumed that swapping fractions are low and government agencies do not reveal the rates to the public because swapping at high levels destroys relationships involving the swapped and unswapped variables.

10.5 Randomization: Adding Random Noise

In randomization, some true information is taken away and some false information is introduced [4]. For example; Numerical data can be protected by adding some randomly selected amount of noise, e.g., a random draw from a normal distribution with mean equal to zero – either to the observed values or to answers to statistical queries. Adding noise to values can reduce the possibilities of accurate matching on

the perturbed data, and distort the values of sensitive variables. The degree of confidentiality protection depends

on the nature of the noise distribution; e.g., using a large variance provides greater protection.

10.6 Synthetic data

The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions. These distributions are specified to reproduce as many of the relationships in the original data as possible.

11. CONCLUSION

This paper has reviewed about Big Data, types of data, different sources of generating data sets, privacy challenges and different privacy techniques that help in preserving data privacy. While recent research has shed much light on formal disclosure metrics and new provably private methods that provide useful statistical information, there are still many intriguing and important research challenges in this area. For instance, most work on privacy has considered data where each record corresponds to a unique individual, and the different records are typically considered independent. Privacy in Big Data has created its impact and this has to be handled carefully in the world where data is everywhere.

12. REFERENCE

[1] Big Data: The End of Privacy or a New Beginning?Ira S.Rubinstein*, October 2012

[2] www.privacyinternnational.org

[3] www.craigbellamy.net

[4] Randomization in Privacy Preserving Data Mining,Alexandre Evfimievski, Cornell University Ithaca, NY14853, USA

[5] Big Data: New Opportunities and New challenges, K.Michael ; Univ. of Wollongong, Wollongong, NSW,Australia ; K. W. Miller, June 2013

[6] An Approach for preserving Privacy in Data Mining, TKachwala, S Parmar, IJARCSSE,2014

[7] Security and Privacy of Big Data, Sithu D. Sudarsan , Raoul P. Jetley,, Srini Ramaswamy, June 2015

[8] A Precautionary Approach to Big Data Privacy,Arvind Narayanan,Joanna Huey, Edward W.Felten,2016

[9] Privacy and Big Data, Masood Mortazavi and Khaled

[10] "Data Mining for Big Data: A Review, Thakur, Bharti,and Manish Mann, 2014