International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Security/Privacy Issues and Challenges in Big Data

Mr. Vishal Joshi

Asst. Professor, Faculty of Computer Applications,
Acropolis Institute of Technology & Research, Indore, Madhya Pradesh, India

---

Abstract: *As always data is one of the most important assets for every Economy, Production, Organization, Business function and individual. The amount of data in world is growing day by day because of use of internet, Smartphone, social network, fine tuning of ubiquitous computing and many other technological advancements. So, a secure framework to social networks is a very hot topic of research. Data captured through various devices, generates ocean of information. Generally size of the data is in Petabyte and Exabyte. The continuous growth in the importance and volume of data has created a new problem: it cannot be handled by traditional analysis techniques. This problem was, therefore, solved through the creation of a new paradigm: Big Data. However, Big Data originated new issues related not only to the volume or the variety of the data, but also to data security and privacy. In addition, the traditional mechanisms to support security such as firewalls and demilitarized zones are not suitable to be applied in computing systems to support Big Data. In this paper, we highlight the some important concept of big data-specific security and privacy challenges so that it will bring renewed focus on fortifying big data infrastructures.*

Keywords: Big data, Security, Privacy,Challenges

1. INTRODUCTION

Big Data is a term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. As far as back in 2001, industry analyst Doung Laney (currently with Gartener), articulated the mainstream of definition of Big Data as the three Vs; Volume, Velocity and Variety. At SAS, SAS considered two additional dimensions when thinking about Big Data: the Variability and Veracity [1]. Oracle defined Big Data in terms of four Vs – Volume, Velocity, Variety and Value [2]. Oguntimilehin define Big Data in terms of five Vs- Volume, Velocity, Variety, Variability, Value and a C-Complexity [3].

2. FEATURES OF BIG DATA

Volume: In the world of big data, when we start talking about volume, we're talking about insanely large amounts of data. Organization collects the data from variety of sources like internet, Smartphone, social network, and many other advancement devices.

Velocity: Velocity refers to the speed at which data is being generated, produced, created, or refreshed. Big data is often available in real-time. Compared to small data, big data are produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing [4].

Variety: Up to 85 percent of an organization's data is unstructured – not numeric – but it still must be folded into quantitative analysis and decision making. Text, video, audio and other unstructured data require different architecture and technologies for analysis [5].

Veracity: Data veracity, in general, is how accurate or truthful a data set may be. In the context of big data, however, it takes on a bit more meaning. More specifically, when it comes to the accuracy of big data, it's not just the quality of the data itself but how trustworthy the data source, type, and processing of it is. Removing things like bias, abnormalities or inconsistencies, duplication, and volatility are just a few aspects that factor into improving the accuracy of big data.

Variability: In addition to the increasing velocities and varieties of data, data flows are unpredictable – changing often and varying greatly. It's challenging, but businesses need to know when something is trending in social media, and how to manage daily, seasonal and event-triggered peak data loads.

Value: Last, but arguably the most important of all, is value. The other characteristics of big data are meaningless if you don't derive business value from the data.

Substantial value can be found in big data, including understanding your customers better, targeting them accordingly, optimizing processes, and improving machine or business performance. You need to understand the potential, along with the more challenging characteristics, before embarking on a big data strategy.

Complexity: Today's data comes from multiple sources and it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out.

3. PRIVACY AND SECURITY ISSUES WHILE USING BIG DATA

3.1 Saving and Retrieving big data

One of the ways to achieve securely storing big data is use of encryption. Once data is encrypted, if the encryption keys are safe, then it is not feasible to retrieve the original data from the encrypted data alone. At the same time, encrypted data must be queried efficiently. Still, the risks of using encrypted data processing [6] need to be further understood to give scalability for the big data as minimizing realistic security and privacy risks. Still if the data is stored in an encrypted format, legal users need to access the data. This shows that we need to have efficient access control techniques that permit users to access the right data. For example, to address new regulations such as right-to-be-forgotten where users may require the deletion of data that belongs to them, we may need to better understand how the data linked and shared among multiple users in a big data system. Study of this area shows understanding how to provide scalable, secure and privacy-aware access control mechanisms for the future big data applications ranging from personalized medicine to Internet of Things systems while satisfying new regulatory requirements.

3.2 Associating and sharing big data

In some cases, data that belongs to diverse sources need to be integrated while satisfying many privacy requirements. To protect individual privacy, only the records belonging to government watch lists may be shared. Obviously, these types of use cases involve connecting potentially sensitive data belonging to the different data controllers. Once data is collected and probably associated/cleaned, it may be shared across organizations to facilitate new applications and release potential value. Therefore, many issues ranging from security to privacy to encouragement for sharing big data need to be considered.

3.3 Studying big data

Another important challenge in using big data is to address the privacy and the security issues in studying big data. Especially, recent developments in machine learning techniques have created important novel applications in many fields ranging from health care to social networking while creating important privacy challenges. Big data processing paradigm categorizes systems into batch, stream, graph, and machine learning processing [7, 8]. 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 unsolicited 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.

3.4 Accountability Issues in big data

As machine learning algorithms change more and more aspects of our lives, it becomes critical to understand how these algorithms transform the way decisions are made in today's data-driven society. The lack of transparency in data-driven decision-making algorithms can easily cover myth and risks codified in the underlying mathematical models, and take care of inequality, bias, and further division between the privileged and the under-privileged [9]. Although the recent research tries to address these transparency challenges [10] more research is needed to ensure fairness, and accountability in usage of machine learning models and big data driven decision algorithms. Understanding the data provenance [11] (i.e., how the data is created, who touched it etc.) have shown to improve trust in decisions and the quality of data used for decision making. In addition to increasing accountability in decision making, more work is needed to make organizations accountable in using privacy sensitive data. With the recent regulations such as GDPR [12]using data only for the purposes consented by the individuals become critical, since personal data can be stored, analyzed and shared as long as the owner of the data consent the data usage purposes. At the same time, it is not clear whether the organizations that collect the privacy sensitive data always process the data according to user consent. An example of this problem is reflected in the recent Cambridge Analytica scandal [13]. In this case, it turns out that the data collected by Facebook is shared for purposes that are not explicitly consented by the individuals which the data belong. As more and more data collected, making organizations accountable for data misuse becomes more critical. It is not clear whether purely technical solutions can solve this problem, even though some research try to formalize purpose based access control and data sharing for big data [14]. Legal and economic solutions (e.g., rewarding insiders that report data misuse) need to be combined with technical solutions. Research that addresses this interdisciplinary area emerges as a critical need.

3.5 Blockchain and Big

The first documented design of blockchain was in 2008, and the first open source implementation of blockchain was deployed in 2009 as an integral element of Bitcoin, the first decentralized digital currency system to distribute bitcoins through the open source release of the Bitcoin peer to peer software. Both were put forward by an anonymous entity,

known as Satoshi Nakamoto [15]. The Bitcoin system uses the blockchain as its distributed public ledger, which records and verifies all bitcoin transactions on the open Bit coin peer to peer networked system. A remarkable innovation of the bitcoin blockchain is its capability to prevent double spending for bitcoin transactions traded in a fully decentralized peer to peer network, with no reliance to any trusted central authority. As a secure ledger, the blockchain organizes the growing list of transaction records into a hierarchically expanding chain of blocks [16] with each block guarded by cryptography techniques to enforce strong integrity of its transaction records. New blocks can only be committed into the global block chain upon their successful competition of the decentralized consensus procedure. The data privacy research in the past decades has shown the risks of privacy leakage due to various inference attacks, which link sensitive transaction data and/or pseudonym to the true identity of the real users, even though only pseudonym is being used [17, 18]. Such privacy leakage can lead to breaching the confidentiality of transaction information. Thus, confidentiality and privacy pose a major challenge for block chain and its applications that involve sensitive transactions and private data.

3.6 Challenger ML and ML for Security

Like many application domains, more and more data are collected for cyber security. Examples of these collected data include system logs, network packet traces, account login formation, etc. Since the amount of data collected is ever increasing, it became impossible to analyze all the collected data manually to detect and prevent attacks. Therefore, data analytics are being applied to large volumes of security monitoring data to detect cyber security incidents [19]. For example, a report from Gartner claims [20] that "Information security is becoming a big data analytics problem, where massive amounts of data will be correlated, analyzed and mined for meaningful patterns." There are many companies that already offer data analytics solutions for this important problem. Of course, data analytics is a means to an end where the ultimate goal is to provide cyber security analysts with prioritized actionable insights derived from big data. Still, direct application of data analytics techniques to the cyber security domain may be misguided. Unlike most other application domains, cyber security applications often face adversaries who actively modify their strategies to launch new and unexpected attacks. The existence of such adversaries in cyber security creates unique challenges compared to other domains where data analytics tools are applied. First, the attack instances are frequently being modified to avoid detection. Hence a future dataset will no longer share the same properties as the current datasets. For example, attackers may change the spam e-mails written by adding some words that are typically associated with legitimate e-mails. Therefore, the spam e-mail characteristics may be changed significantly by the spammers as often as they want. Secondly, when a previously unknown attack appears, data analytics techniques need to respond to the new attack quickly and cheaply. For example, when a new type of ransomware appears in the wild, we may need to update existing data analytics techniques quickly to detect such attacks.

4. CHALLENGES IN USING BIG DATA

Challenges are always there to cope up with grabbing opportunities. To handle these challenges, we need to know various computational complexities, security threats, and computational techniques of big data to analyze big data problems. For instance, the mathematical and statistical methods that work well for small data set do not work well with large data sets. Likewise, many computational methods that work well for small data won't work well with big data. The challenges of using big data are as follows.

4.1 Insufficient understanding and acceptance of big data

Many times, companies fail to know even the basics: what big data is, what its application are, what setup is needed, etc. Without a clear understanding, a big data adoption project risks to be ruined to failure. Companies may waste lots of time and resources on things they don't even know how to use [21].

Solution:

Big data, being a huge change for a company, should be accepted by top management first and then down the ladder. To ensure big data acceptance at all levels, Information Technology department need to organize various trainings and workshops to get acquainted with all the possible flavor of big data.

To see to big data acceptance even more, the implementation and use of the new big data solution need to be monitored and controlled. However, top management should not overdo with control because it may have an adverse effect.

4.2 Availability of vast big data technologies

Today's it is very easy to get lost in the variety of big data technologies now prevailing in the market. Do you need Spark or would the speeds of Hadoop, MapReduce be enough? Is it better to store data in Cassandra or HBase? Finding the answers can be tricky. And it's even easier to

choose poorly, if you are exploring the vast of opportunities from technological perspective without a clear view of what you need.

Solution:

As big data is new, trying to seek expert help would be the right way to proceed. You could hire an expert or turn to a vendor for big data consulting. In both cases, with collaborative efforts, you will be able to work out on planned basis, afterwards, choose the needed technology stack.

4.3 Incurring cost is very high

Big data implementation projects require lots of operating expense. If you choose for an on-premises solution, you will have to pay the costs of new hardware, new hires (administrators and developers), and electricity and so on. Moreover: although the needed frameworks are open-source, you will have to pay for the development, setup, configuration and maintenance of new software.

If you choose on a cloud-based big data solution, you will still want to appoint staff (as above) and pay for cloud services, big data solution development as well as setup and maintenance of needed frameworks.

Solution:

The particular recovery of your company's wallet will depend on your company's specific technological requirements and industry goals. For instance, companies who want flexibility benefit prefer cloud. Whereas companies with extremely insensitive security requirements go on-premises.

There are also hybrid solutions available, in which parts of data are stored and processed at cloud and parts on-premises, which can also be cost-effective. Moreover, resorting to data lakes or algorithm optimizations (if done properly) can also save money. All in all, the key to solve these challenges is to correctly analyze your needs and choose a corresponding course of action.

4.4 Difficulty in handling quality data

At any stage, problem of data integration occur, since the data you want to analyze comes of different sources in variety of different formats. For example, e-commerce firms need to analyze data from call centers, competitor's websites, website logs and social media. Data formats naturally differ and matching them can be problematic.

Fact is that bit data is not 100% accurate. So you need to control its inaccuracy to make it authentic up to some extent. Big data may contain wrong, duplicate, and contradictory information. It may be possible that inferior quality data may bring insight of bright opportunities to your business task.

Solution:

There are several techniques available for cleansing data. Your big data needs to have a proper model. Only after creating that, you can go ahead and do other things.

- Compare data to the single point of truth (for instance, compare variants of addresses to their spellings in the postal system database).
- Match records and merge them, if they relate to the same entity.

Remember that big data is never 100% accurate. You have to deal with it.

4.5 Complex process of converting big data into valuable insights

It require for organizations to invest resources and executive attention towards some of the key Big Data challenges such as identifying and securing the right mix of skills and capabilities to turn data into value. In fact, according to analysts, businesses are now facing a yawning gap between demand and capability that – in the US alone – will require some 1.5m data-savvy managers and analysts to fill. In other words, those that hope to convert their data into business value must first start by identifying and securing sufficient and experienced resources to dig out from the data avalanche [22].

Solution:

The reason that you failed to have the needed items in stock is that your big data tool doesn't analyze data from social networks or competitor's web stores. While your rival's big data among other things does note trends in social media in near-real time. And their shop has both items and even offers a 15% discount if you buy both.

The idea here is that you need to create a proper system of factors and data sources, whose analysis will bring the needed insights, and ensure that nothing falls out of scope. Such a system should often include external sources, even if it may be difficult to obtain and analyze external data.

4.6 Difficulty in ranging up

The most common feature of big data is its striking ability to grow up. This is one of the most serious testing's of big data is associated closely with this.

Your solution's design may be thought through and adjusted to up scaling with no extra efforts. But the real problem is not the actual process of bring in new processing and storing abilities. It lies in the difficulty of scaling up so, that your system's performance doesn't turn down and you stay within budget.

Solution:

The first and primary safeguard for challenges like this is a well-brought-up design of your big data solution. As long as your big data solution can possess such a thing, fewer troubles are likely to occur later. Another most important thing to do is designing your big data algorithms while keeping future up scaling in mind.

But besides that, you also need to plan for your system's maintenance and support so that any changes related to data growth are properly attended to. And on top of that, holding systematic performance audits can help identify weak spots and timely address them.

5. CONCLUSION

This paper presents the fundamental concepts of Big Data. These concepts include the increase in data in various organizations, and the role of Big Data in the current environment of enterprise and technology. This paper provide explanation of how various users of big data faces issues in dealing with day to day operations in big data at various stages of big data ecosystem. This paper gives an explanation of the research took place in order to address the main problems and challenges related to security in Big Data, and thrown light on points of consideration while working with big data. This paper reveals that although the information security standards, methodologies and S/W to ensure the security and privacy of the Big Data environment already exist, the few meticulous characteristics of Big Data make them futile if they are not used in an integrated manner. This paper also presents some research for these challenges, but it does not provide a concrete solution for the problem. Although it points to some information and technologies that may add the most relevant and challenging Big Data security and privacy issues. This paper also includes various challenges faced by various professional workers in different domains. These challenges are generic and should be considered on high priority before undertaking any project/assignment. Each emphasized challenge also includes the solution to the challenges that should be executed before going for big data. Paper also includes challenges which can be unfolded in view of initial phase of any big project. In conclusion, the Big Data technology acquiring hold in various industries, and that is the reason why there have been a number of studies created in last few years. It is continuous process of studying all above mentioned issues in deep, to come out with concrete solutions, in fact, the studies created from now should focus on more specific problems.

REFERENCES

[1] https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[2] Oracle (2013), Information Management and Big Data: A Reference Architecture, www.oracle.com/.../info-mgmt-big-data-r..., retrieved 20/03/14.

[3] A Review of Big Data Management, Benefits and Challenges 1 Oguntimilehin A., 2 Ademola E.O. 1,2 Department of Computer Science, Afe Babalola University, Ado-Ekiti, Nigeria.

[4] "Data, data everywhere"(https://www.economist.com/special-report/2010/02/27/data-data-everywhere). The Economist. 25 February 2010. Retrieved 9 December 2012.

[5] (http://eric.univ-lyon2.fr/~ricco/cours/slides/sources/big-data-meets-big-data-analytics-105777.pdf)

[6] Islam, M. S., Kuzu, M., and Kantarcioglu, M. (2012). "Access pattern disclosure on searchable encryption: Ramification, attack and mitigation," in 19th Annual Network and Distributed System Security Symposium, NDSS 2012 (San Diego, CA).

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

[8] Zhang Y, Cao T, Li S, Tian X, Yuan L, Jia H, Vasilakos AV. Parallel processing systems for big data: a survey. In: Proceedings of the IEEE. 2016.

[9] Sweeney, L. (2013). Discrimination in online ad delivery. Commun. ACM 56, 44–54. doi: 10.1145/2447976.2447990.

[10] Baeza-Yates, R. (2018). Bias on the web. Commun. ACM 61, 54–61. doi: 10.1145/3209581

[11] Bertino, E., and Kantarcioglu, M. (2017). "A cyber-provenance infrastructure for sensor-based data-intensive applications," in 2017 IEEE International Conference on Information Reuse and Integration, IRI

2017 (San Diego, CA), 108–114. doi: 10.1109/IRI.2017.91

[12] Voigt, P., and Bussche, A. V. D. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer Publishing Company, Incorporated.

[13] Cadwalladr, C., and Graham-Harrison, E. (2018). Revealed: 50 million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach. Available online at: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (Accessed on 12/21/2018).

[14] Ulusoy, H., Kantarcioglu, M., Pattuk, E., and Kagal, L. (2015b). "Accountablemr: toward accountable mapreduce systems," in 2015 IEEE International Conference on Big Data, Big Data 2015 (Santa Clara, CA), 451–460. doi: 10.1109/BigData.2015.7363786.

[15] Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system. www.Bitcoin.Org, 9. (2008).

[16] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. 2016. Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction.

[17] Jules DuPont and Anna Cinzia Squicciarini. [n. d.]. Toward De-Anonymizing Bitcoin by Mapping Users Location. In CODASPY 2015. 139–141.

[18] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. [n. d.]. A Fistful of Bitcoins: Characterizing Payments Among Men with No Names. In IMC 2013. 127–140.

[19] Shaon, F., and Kantarcioglu, M. (2016). "A practical framework for executing complex queries over encrypted multimedia data," in Proceedings on 30th Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy XXX DBSec 2016 (Trento), 179–195. doi: 10.1007/978-3-319-41483-6_14.

[20] MacDonald, N. (2012). Information Security is Becoming a Big Data Analytics Problem. Available online at: https://www.gartner.com/doc/1960615/information-security-big-data-analytics (Accessed Jul 15, 2018).

[21] https://www.scnsoft.com/blog/big-data-challenges-and-their-solutions.

[22] https://www.ft.com/content/6bcc5a9c-37f1-11e2-b8d3-00144feabdc0.