International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 13, Issue 3 (March 2017), PP.18-24

An Investigation on Scalable and Efficient Privacy Preserving Challenges for Big Data Security

Mr. S. Dhinakaran1, Dr. J. Thirumaran[2]
Research Scholar, Department Of Computer Science, Rathinam College Of Arts & Science, Coimbatore.

ABSTRACT:- Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making. Big data refers to huge amount of digital information collected from multiple and different sources. With the development of application of Internet/Mobile Internet, social networks, Internet of Things, big data has become the hot topic of research across the world, at the same time; big data faces security risks and privacy protection during collecting, storing, analyzing and utilizing. Since a key point of big data is to access data from multiple and different domains security and privacy will play an important role in big data research and technology. Traditional security mechanisms, which are used to secure small scale static data, are inadequate. So the question is which security and privacy technology is adequate for efficient access to big data.

This paper introduces the functions of big data, and the security threat faced by big data, then proposes the technology to solve the security threat, finally, discusses the applications of big data in information security. Main expectation from the focused challenges is that it will bring a novel focus on the big data infrastructure.

Keywords:- Big Data, Security, Challenges, Issues, Privacy.

I. INTRODUCTION

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabyte (1015 or 1000 terabytes per petabyte) as big data.

The complex nature of big data is primarily driven by the unstructured nature of much of the data that is generated by modern technologies, such as that from web logs, radio frequency Id (RFID), sensors embedded in devices, machinery, vehicles, Internet searches, social networks such as Facebook, portable computers, smart phones and other cell phones, GPS devices, and call center records. In most cases, in order to effectively utilize big data, it must be combined with structured data (typically from a relational database) from a more conventional business application, such as Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM).

Similar to the complexity, or variability, aspect of big data, its rate of growth, or velocity aspect, is largely due to the ubiquitous nature of modern on-line, real-time data capture devices, systems, and networks. It is expected that the rate of growth of big data will continue to increase for the foreseeable future. Specific new big data technologies and tools have been and continue to be developed. Much of the new big data technology relies heavily on massively parallel processing (MPP) databases, which can concurrently distribute the processing of very large sets of data across many servers. As another example, specific database query tools have been developed for working with the massive amounts of unstructured data that are being generated in big data environments.

The purpose of this article, therefore, is to sketch the emergence of Big Data as a research topic from several points: (1) timeline, (2) geographic output, (3) disciplinary output, (4) types of published papers, and (5) thematic and conceptual development. The amount of data available to us is increasing in manifold with each passing moment. Data is generated in huge amounts all around us. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. [1] With the advancement in technology, this data is being recorded and meaningful value is being extracted from it. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.

The 3Vs that define Big Data are Variety, Velocity and Volume.

1) Volume: There has been an exponential growth in the volume of data that is being dealt with. Data is not just in the form of text data, but also in the form of videos, music and large image files. Data is now stored in terms of Terabytes and even Petabyte in different enterprises. With the growth of the database, we need to re-evaluate the architecture and applications built to handle the data.

2) Velocity: Data is streaming in at 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. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

3) Variety: Today, data comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. We need to find ways of governing, merging and managing these diverse forms of data.

There are two other metrics of defining Big Data

4) Variability: Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

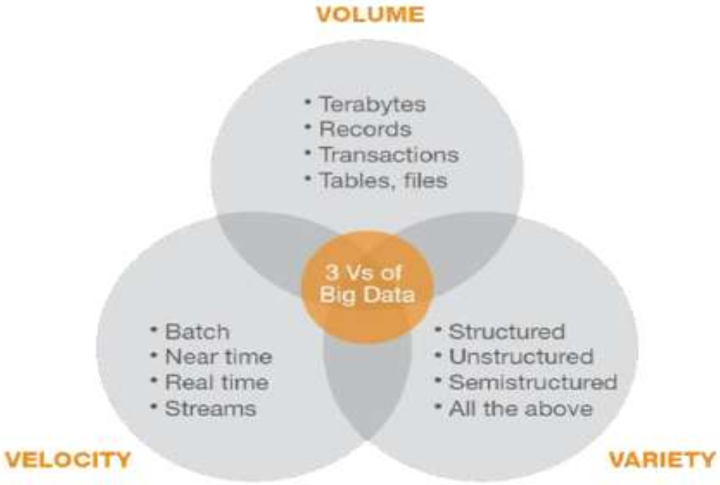

Fig 1.1: - 3V's of Big Data

With the deepening of Internet applications, social networks and internet of things produced a huge amount of data, which we called big data. It makes the analysis and application of the data more complex, and difficult to manage. These data, including text, images, audio, video, Web pages, e-mail, micro blogging and other types, Among them, 20% are structured data, 80% are semi-structured and unstructured data. big data is large and complex, so it is difficult to deal with the existing database management tools or data processing application. Commercial enterprise collect information on all aspects of customers for a long time, to analyze the user behavior law, more accurately portray the individual characteristics, to provide users with better personalized products and services, and more accurate advertising recommended.

Using Big Data, security functions are required to work over the heterogeneous composition of diverse hardware, operating systems, and network domains.

II. UNDERSTANDING BIG DATA

Big data has many sources. For example, every mouse click on a *web site* can be captured in Web log files and analyzed in order to better understand shoppers' buying behaviors and to influence their shopping by dynamically recommending products. *Social media* sources such as Facebook and Twitter generate tremendous amounts of comments and tweets. This data can be captured and analyzed to understand, for example, what people think about new product introductions. *Machines*, such as smart meters, generate data. These meters continuously stream data about electricity, water, or gas consumption that can be shared with customers and combined with pricing plans to motivate customers to move some of their energy consumption, such as for washing clothes, to non-peak hours.

There is a tremendous amount of *geospatial* (e.g., GPS) data, such as that created by cell phones, that can be used by applications like Four Square to help you know the locations of friends and to receive offers from nearby stores and restaurants. *Image, voice, and audio* data can be analyzed for applications such as facial recognition systems in security systems.

During the last few years, big data has evolved to an emerging field where innovation on technology allows for new ways to deal with huge amounts of data created in near real time by a vast variety of sources (IoT sensing devices, M2M communication, social applications, mobile video, etc.). Big data can provide for "big" analytics that offer novel opportunities to reuse and extract value from the "information chaos", escaping the confines of structured databases, identifying correlations, conceiving new, unanticipated uses of data. Big analytics can offer a whole new area of opportunities from research to online transactions and service provision in several sectors of everyday life. This has been recognized by the European Commission, which in its latest Communication on big data stresses the need for a data-driven economy, contributing to citizens' welfare and socio-economic growth [1].

Business consultants Gartner Inc. define big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [7]. This definition points out the three most outlined dimensions of big data (also known as the 3Vs3 that define big data):

- **Volume**: Huge amounts of data in the scale of Zettabytes 4 and more.
- **Velocity**: Real time streams of data flowing from diverse resources (e.g. physical sensors or "virtual sensors" from social media, such as Twitter streams).
- **Variety**: Data from a vast range of systems and sensors, in different formats and datatypes.

For the purpose of this report one more interesting dimension of big data is also veracity, which describes the incompleteness (inconsistency, inaccuracy) of data.

Due to the above-mentioned characteristics, big data is seen today as the new opportunity for analytics that can offer significant advancements to several aspects of our everyday life, including health, leisure, environment, employment, etc. To this end, data has been characterized by many as the fuel of the 21st century economy or the new oil6. Still, data differ from oil in one critical element: they are not just an available (and rather difficult to find) resource but, on the contrary, they are constantly generated by people's activities. This is why big data may (and in many cases do) also involve personal data, for example a name, a picture, contact details, posts on social networking websites, healthcare data, location data or a computer IP address.

2.1. Big Data Analytics

The term "big data analytics" refers to the whole data management lifecycle of collecting, organizing and analyzing data to discover patterns, to infer situations or states, to predict and to understand behaviours. Its value chain includes a number of phases that can be summarized as follows:

- **Data Acquisition/Collection:** the process of gathering, filtering and cleaning data before it is put in a data repository or any other storage solution on which data analysis can be carried out. Examples of potential sources are social networks, mobile apps, wearable devices, smart grids, online retail services, public registers, etc. As the main purpose is to maximize the amount of available data (so as to appropriately feed the analysis), the process is usually based on fast and massive data collection, thus, assuming high-volume, high-velocity, high-variety, and high-veracity but low-value data.
- **Data Analysis:** the process concerned with making the "raw" collected data amenable for decision-making as well as domain specific usage. The key challenge of data analysis is to find what is really useful. A critical element in that respect is to combine data from different sources in order to derive information that cannot be found otherwise. Data analysis covers structured or unstructured data, with/without semantic information and can have multiple levels of processing and different techniques (e.g. diagnostic, descriptive, predictive and prescriptive analysis).
- **Data Curation:** the active management of data over its lifecycle to ensure it meets the necessary quality requirements for effective usage. It includes functions like content creation, selection, classification, transformation, validation and preservation. A main aspect in that respect is the need to assure the reusability of the data, not only within their original context but in many different contexts.
- **Data Storage:** storing and managing data in a scalable way satisfying the needs of applications/analytics that require access to the data. Cloud storage is the trend but in many cases distributed storage solutions would be best options (e.g. for streaming data).
- **Data Usage:** covers the use of the data by interested parties and is very much dependent on the data processing scenario. For example the results from an analysis on trends in mobile apps usage could be available for the general public or restricted to a mobile service provider who commissioned the study. Therefore, users of big data may vary from a single organisation to a wide range of parties, such as banks, retailers, advertising networks, public authorities, etc.

A variety of stakeholders are involved in the different phases of the big data value chain, including hardware, software and operating system vendors, different types of service providers (telecom operators, social networks, cloud providers, etc.), analytics providers, data brokers, public authorities, etc. These stakeholders may take different roles during a big data analytics scenario and interact with each other in variant ways.

III. BIG DATA SECURITY & PRIVACY

Privacy and security in terms of big data is an important issue. Big data security model is not suggested in the event of complex applications due to which it gets disabled by default. However, in its absence, data can always be compromised easily. As such, this section focuses on the privacy and security issues.

- **Privacy** Information privacy is the privilege to have some control over how the personal information is collected and used. Information privacy is the capacity of an individual or group to stop information about them from becoming known to people other than those they give the information to. One serious user privacy issue is the identification of personal information during transmission over the Internet [13].
- **Security** Security is the practice of defending information and information assets through the use of technology, processes and training from:-Unauthorized access, Disclosure, Disruption, Modification, Inspection, Recording, and Destruction.

Data privacy is focused on the use and governance of individual data—things like setting up policies in place to ensure that consumers' personal information is being collected, shared and utilized in appropriate ways. Security concentrates more on protecting data from malicious attacks and the misuse of stolen data for profit [14]. While security is fundamental for protecting data, it's not sufficient for addressing privacy. Table 3.1 focuses on additional difference between privacy and security.

S. No	Privacy	Security
1.	Privacy is the appropriate use of user's information	Security is the "confidentiality, integrity and availability" of data
2.	Privacy is the ability to decide what information of an individual goes where	Security offers the ability to be confident that decisions are respected
3.	The issue of privacy is one that often applies to a consumer's right to safeguard their information from any other parties	Security may provide for confidentiality. The overall goal of most security system is to protect an enterprise or agency
4.	It is possible to have poor privacy and good security practices	However, it is difficult to have good privacy practices without a good data security program

Table 3.1: - Privacy Vs Security

Security and privacy issues are magnified by velocity, volume and variety of big data, such as large scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition ,and high volume inter-cloud migration. The use of large scale cloud infrastructure with diversity of software platforms, spread across large networks of computers, also increases the attack surface of entire system. therefore traditional security mechanisms, which are tailored to securing small scale static(as opposed to streaming)data, are inadequate. Ex. analytics for anomaly detection would generate too many outliers.

Analyzing data in order to support decision making, discover trends or open new business opportunities is not new. Neither is the collision of certain types of such processing with privacy and data protection principles.

The scale is in terms of volume, variety, velocity and veracity, all the Vs of the big data definition, and their combination in analytics technologies. To this end, the main privacy challenges associated with big data, overrunning the privacy concerns of "traditional" data processing, are as follows:

- **Lack of control and transparency**
  The collection of big data is based on so many different and unexpected sources that control by the individual can easily be lost, as in many cases he/she is not even aware of the processing or cannot trace how data are flowing from one system to another. This poses significant engineering challenges on how to effectively and timely inform users and on who is in charge of this task, especially when the processing requires the interaction of many players.
- **Data reusability**
  One of the main targets of big data analytics is to use data, alone or in combination with other data sets, beyond their original point and scope of collection. The scalability of storage allows for potential infinite space, which means that data can be collected continuously until a new value can be created from insights derived out of them.
- **Data inference and re-identification**
  Another important element for big data is the possibility to combine data sets from many different sources, so as to derive more (and new) information. This also triggers privacy risks, especially if linking different sources may allow the emergence of patterns related to single individuals.

- **Profiling and automated decision making**

The analytics applied to combined data sets aim at building specific profiles for individuals that can be used in the context of automated decision making systems, e.g. for offering or excluding from specific services and products.

On top of the risks mentioned above, one additional challenge in big data is the difficulty in enforcing, and/or monitoring the data protection controls. These difficulties are similar to the ones of cloud computing, for example in relation to the location of the data and the possibility to conduct audits. Still, in big data there is one additional and critical element: the involvement and interaction of many diverse stakeholders, which makes it even more difficult (for regulators, data controllers and users) to identify privacy flaws and impose relevant measures.

Therefore, the new thing in big data is not the analytics itself or the processing of personal data. It is rather the new, overwhelming and increasing possibilities of the technology in applying advanced types of analyses to huge amounts of continuously produced data of diverse nature and from diverse sources. The data protection principles are the same. But the privacy challenges follow the scale of big data and grow together with the technological capabilities of the analytics.

IV. PRIVACY ENHANCING TECHNIQUES IN BIG DATA

There is no single magical solution to solve the identified Big Data security and privacy challenges and traditional security solutions, which are mainly dedicated to protect small amounts of static data, are not adequated to the novel requisites imposed by Big Data services (Cloud Security Alliance, 2013). There is the need to understand how the collection of large amounts of complex structured and unstructured data can be protected. Non-authorized access to that data to create new relations, combine different data sources and make it available to malicious users is a serious risk for Big Data. The basic and more common solution for this includes encrypting everything to make data secure regardless where the data resides (data center, computer, mobile device, or any other). As Big Data grows and its processing gets faster, then encryption, masking and tokenization are critical elements for protecting sensitive data.

We provide an overview of key identified technologies that could be further applied and/or explored in the particular case of big data. Having said that, it is important to note that most of these technologies are available today and have been used in the “traditional” processing of personal data [6]. But again it is all about the scale: in the context of this report we aim to explain the specificities and adoption of certain technologies in big analytics, taking account the volume, velocity, variety and veracity of the new data landscape.

In particular, we first present anonymization, which has been so far the “traditional” approach towards data analytics, facing some new challenges in the era of big data. Then we explore the developments in cryptography and more specifically encrypted search, which can allow for privacy preserving analytics without disclosing personal data. On top of these approaches, we discuss privacy by security, i.e. ensuring an overall security framework for the protection of personal data, especially regarding access control policies and their enforcement. Transparency and control mechanisms are also central in big data, in order to offer information and choice to the individuals. To this end, we examine notice, consent and other mechanisms relying on users’ privacy preferences and taking into account relevant usability issues.

4.1. Data Privacy Protection Technology

To protect the privacy of big data, even if the data with privacy leak, the attacker can’t obtain the effective value of data. We can use data encryption and Data anonymity technology

(1) Data Encryption Technology Data encryption technology is an important means to protect data confidentiality, it safeguards the confidentiality of the data, but it cut down the performance of the system at the same time. The data processing ability of big data system is fast and efficient, which can satisfy the requirements of the hardware and software required for encryption. So the homomorphic encryption has become a research hotspot in data privacy protection. The homomorphic encryption is a model for the calculation of the cipher text, avoiding the encryption and decryption in the unreliable environment, and directly operation on the cipher text. Which is equivalent to the procedure of processing the data after decryption, then encrypting it? Homomorphic encryption is still in the exploratory stage, the algorithm is immature, low efficiency, and there is a certain distance away from practical application.

(2) Data Anonymity Technology Data anonymity is another important technology for privacy protection, Even if the attacker gets the data that contains the privacy, he can't get the original exact data, because the value of the key field is hidden. However, in the background of big data, the attacker can obtain data from multiple sources, then associate the data from one source with another source, then will find the original meaning of the hidden data.

(3) Generalization Technology The third technology of privacy protection is generalization technology, which is to generalize the original data, so that the data is fuzzy, so as to achieve the purpose of privacy protection.

4.2. Access Control Technology

Big data contains a wealth of information resources, all professions and trades have great demand of the data, so we must manage access rights of big data carefully. Access control is an effective means to achieve controlled sharing of data, but in big data environment, the number of users is huge, the authority is complex, and a new technology must be adopted to realize the controlled sharing of data.

(1) Role Mining Role-based access control (RBAC) is an access control model used widely. By assigning roles to users, roles related to permissions set, to achieve user authorization, to simplify rights management, in order to achieve privacy protection. In the early, RBAC rights management applied "top-down" mode: According to the enterprise's position to establish roles, When applied to big data scene, the researchers began to focus on "bottom-up" mode, that is based on the existing "Users - Object" authorization, design algorithms automatically extract and optimization of roles, called role mining. In the big data scene, using role mining techniques, roles can be automatically generated based on the user's access records, efficiently provide personalized data services for mass users. It can also be used to detect potentially dangerous that user’s behavior deviates from the daily behavior. But role mining technology are based on the exact, closed data set, when applied to big data scene, we need to solve the special problems: the dynamic changes and the quality of the data set is not higher.

(2) Risk Adaptive Access Control In big data scene, the security administrator may lack sufficient expertise, Unable to accurately specify the data which users can access, risk adaptive access control is an access control method for this scenario. By using statistical methods and information theory, define Quantization algorithm, to achieve a risk-based access control. At the same time, in big data environment, to define and quantify the risk are more difficult.

4.3. Data Provence technology

Due to the diversification of data sources, it is necessary to record the origin and the process of dissemination and calculation, provide additional support for the latter mining and decision. Before the emergence of the concept of big data, Data Provence technology has been widely studied in database fields. Its purpose is to help people determine the source of the data in the data warehouse. The method of data Provence is labeled method through the label, we can know which data in the table is the source, and can easily checking the correctness of the result, or update the data with a minimum price. In the future, data Provence technology will play an important role in the field of information security. But data provence technology for big data security and privacy protection also need to solve the following two questions: 1, The balance between privacy protection and data provence;2, to protect the security of data provence technology itself.

4.4. Consent, ownership and control

User control is a crucial goal in big data and it can be reached through a multichannel approach. Consent is one possible solution (and probably the most prominent one). Other methods and tools can also contribute by ensuring accurate audits and determining the compliance of controllers and processors with the rules. An example of such a method is “tagging” every unit of personal data with “metadata” describing data protection requirements. This is also the perspective of semantic web, but putting tags and rules on data is a costly activity which will require a multi-stakeholder effort.

Practical implementation of consent in big data should go beyond the existing models and provide more automation, both in the collection and withdrawal of consent. Software agents providing consent on user’s behalf based on the properties of certain applications could be a topic to explore. Moreover, taking into account the sensors and smart devices in big data, other types of usable and practical user positive actions, which could constitute consent (e.g. gesture, spatial patterns, behavioral patterns, motions), need to be analyzed.

V. CONCLUSIONS

In this paper, we have focused on the security and privacy problems that need to provide more secure for Big Data processing and computing infrastructure. Common elements of Big Data arise from the use of multiple infrastructure tiers for processing Big Data. Big data [2] is analyzed for bits of knowledge that leads to better decisions and strategic moves for overpowering businesses. Yet only a small percentage of data is actually analyzed. In this paper, we have investigated the privacy challenges in big data by first identifying big data privacy requirements and then discussing whether existing privacy preserving techniques are sufficient for big data processing. Privacy challenges in each phase of big data life cycle [7] are presented along with the advantages and disadvantages of existing privacy-preserving technologies in the context of big data applications. This paper also presents traditional as well as recent techniques of privacy preserving in big data.

REFERENCES

[1]. http://www.internetlivestats.com/twitter-statistics/

[2]. http://www.internetlivestats.com/google-search-statistics/

[3]. Grand Challenge: Applying Regulatory Science and Big Data to Improve Medical Device Innovation, Arthur G. Erdman∗, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, NO. 3, MARCH 2013

[4]. http://lsst.org/lsst/google

[5]. http://www.economist.com/node/15557443

[6]. Dona Sarkar, Asoke Nath, “Big Data – A Pilot Study on Scope and Challenges”, International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS, ISSN: 2371-7782), Volume 2, Issue 12, Dec 31, Page: 9-19(2014).

[7]. Sagiroglu, S.; Sinanc, D. ,”Big Data: A Review”
Grosso, P. ; de Laat, C. ; Membrey, P.,(” Addressing big data issues in Scientific DataInfrastructure”

[8]. Kogge, P.M.,(20-24 May,2013), “Big data, deep data, and the effect of system architectures on performance” Szczuka, Marcin,(24-28 June,2013),” How deep data becomes big data”.

[9]. META Group. "3D Data Management: Controlling Data Volume, Velocity, and Variety." February 2001. Performance Analysis of Data Encryption Algorithms: Abdel-Karim Al Tamimi

[10]. FENG Deng-Guo, ZHANG Min, LI Hao. Big Data Security and Privacy Protection[J]. Chinese Journal of Computers, 2014,37(1):246-258.

[11]. MA Li-chuan, PEI Qing-qi, LENG Hao, LI Hong-ning. Survey of Security Issues in Big Data[J]. Ｒadio Communications Technology, 2015,41(1):1-7.

[12]. Hu Kun, Liu Di, Liu Minghui. Research on Security Connotation and Response Strategies for Big Data[J]. Telecommunications Science, 2014(2):112-117,122.

[13]. WANG Yu-long，ZENG Meng-qi. Big Data Security based on Hadoop Architecture[J]. Information Security and Communications Privacy, 2014(7):83-86.

[14]. Big Data Working Group. Big Data Analytics for Security Intelligence[EB/OL]. https://www.cloudsecurityalliance.org/research/big-data.

[15]. Guillermo Lafuente. The big data security challenge[J]. Network Security,2015.(1):12-14.

[16]. Hrestak D, Picek S. Homomorphic Encryption in the Cloud［C］‖2014 37th International Convention on Infor-mation and Communication Technology，Electronics and Micro electronics(MIPＲO)，2014: 1400-1404.

[17]. L.D. Cohen. NOTE On Active Contour Models and Balloons. Computer Vision and Image Processing: Image Understanding, 53:211-218, March 1991.