Big Data Security – The Big Challenge

Minit Arora, Dr Himanshu Bahuguna

Abstract— In this paper we discuss the issues related to Big Data. Big Data is the vouluminous amount of data with variety in its nature along with the complexity of handling such data. In addition to the problem of mining information from Big Data, privacy is a big challenge for big data. In this paper we discuss the issues concerning privacy of this data and some of the existing techniques to ensure privacy of the data.

ms— Big Data, Big Data Analytics, Cloud computing, Big Data Privacy, Privacy requirements inr big data, Big Data security techniques, Encryypted data, De-identification.

1 INTRODUCTION

THE term "Big Data" [1] is used to define massive volume data, both structured and unstructured in nature. The enormous volume of data makes it impossible to process using traditional database and software technologies. It requires —massively parallel software running on tens, hundreds, or even thousands ofservers

1.1 Characteristics of Big Data

Big Data is characterized by the 3 V's- Volume, Variety and Velocity [2]

Volume – It refers to the quantity of data that is generated. It is the size of the data whichdetermines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not.The large scale and rise of size makes it difficult to store and analyze using traditional tools.

Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to. Data comes in all types of formats emails, video, audio, transactions etc.,

.

Velocity - The term 'velocity' refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.This means how fast the data is being produced and how fast the data needs to be processed tomeet the demand.dimensions are also important defining characteristics of Big Data

Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Variability - This is a factor which can be a problem for those who analyze the data. This refers to the inconsistencywhich can be shown by the data at times, thus hampering the process of being able to handle and manage the dataeffectively.

Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data

An

3 BIG DATA ANALYTICS

Big data analytics is the process of examining large data sets containing a variety of data types in big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information .The analytical findings can lead to more effective marketing, new revenue opportunities, better custom service,improved operational efficiency, competitive advantages over rival organizations and other business benefits.The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analyticsprofessionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI)programs. That could include Web server logs social media content andsocial network activity reports, from customer emails and survey responses, mobile-phone call detail records Big data can include both structured and unstructured data.

Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually for example, real-time data on the performance of mobile applications

3 HOW BIG DATA RELATES TO CLOUD COMPUTING

In the cloud computing context, network-accessible resources are de-

- *Minit Arora is currently Assistant Professor, SGRR Institute of Technology and Science, Dehradu, Indian. Ph-,9897593174. E-mail: auminitaro-ra@yahoo.com*
- *Dr Himanshu Bahugunae is currently Professor, Shivalik College of Engineering, Dehradun. India.*

fined as services. These services are typically delivered via one of three cloud computing service models:

1. Infrastructure as a service (IaaS) offers storage, computation, and network capabilities to service subscribers through virtual machines (VMs).

2 Platform as a service (PaaS) provides an environment for software application development and hosts a client's applications in a PaaS provider's computing infrastructure.

3 Software as a service (SaaS) delivers on-demand software services via a computer network, eliminating the cost of purchasing and maintaining software.

Big data analytics use computation intensive data mining algorithms that require efficient high performance processors to produce timely results. Cloud computing infrastructures can serve as an effective platform for addressing both the computational and data storage needs of big data analytics applications. Much big data already resides in the cloud, and this trend will increase in the future. For example,IT research and advisory firm Gartner estimates that, by 2016, more than half of large companies' data will be stored in the cloudThis trend requires that clouds becomethe infrastructure for implementing pervasive and scalable data analytics platforms Coping with and gaining valuefrom cloud-based big data requires novel software tools and innovative analytics techniques.

These technical and business advantages come at a cost. The security vulnerabilities inherited from the underlying technologies (that is, virtualization, IP, APIs, and datacenter) prevent organizations from adopting the cloud in many critical business applications

.4 CLOUD BASED DATA ANALYTICS

Big data refers to massive, heterogeneous, and often unstructured digital content that is difficult to process using traditionaldata management tools and techniques. The term encompasses the complexity and variety of data and data types, real-time datacollection and processing needs, and the value that can be obtained by smart analytics Advanced data mining techniquesand associated tools can help extract information from large, complex datasets that is useful in making informed decisionsin many business and scientific applications including tax payment collection, market sales, social studies, biosciences, and high energy physics. Combining big data analytics and knowledge discovery techniques with scalable computing systems will produce new insights in a shorter time

Although few cloud-based analytics platforms are available today, current research work anticipates that they will becomecommon within a few years.

Some current solutions are based on source systems such as Apache Hadoop and SciDB, while others are proprietary solutions provided by companies such as Google, IBM,EMC, BigML, Splunk Storm, KognitioandInsightsOne

5 BIG DATA PRIVACY AND SECURITY- BIG CHALLENGE

Big Data Security is a big challenge due to the following vulnerabilities

1. Big Data increases the risk ofinformation leakage due to its high volume and velocity.

2. Development of intelligent terminals increases the risk that relates to privacy and prediction of people's behavior

5.1 Privacy Requirements

Probably the most challenging and concerned problem in Big Data is security and privacy. Governmental agencies, the health care industry, biomedical researchers, and private businesses invest enormous resources into the collection, aggregation, and sharing of large amounts of personal data for the enormous benefit of Big Data. Through recent disclosure, the National Security Administration routinely collects and analyzes massive amounts of personal data derived from heterogeneous data sources such as telecommunications, the Internet, and the user databases of large businesses, including Microsoft, Yahoo, Google, Facebook, PalTalk, YouTube, Skype, AOL, and Apple" Many facts show that Big Data will harm the user's privacy if it is not properly handled. The security and privacy issues which should be concerned in Big Data context include:

The personal information of a person when combined with external large data sets leads to the inference of new facts about that person and it's possible that these kinds of facts about the person that are secretive and the person might not want the Data Owner to know or any person to know about them;

1. Information regarding the users (people) is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of;

2. Another important consequence arising would be Social stratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse;

3. Big Data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated

4. The field of privacy in big data which contains a bunch of challenges involves interaction with individuals, re-identification attacks, probable and provable results, and economic effects. Interaction with individuals includes providing transparency, getting consent, revocation of consent and deletion of personal data. Re-identification attacks which have three sub-categories named correlation attacks, arbitrary identification attacks, and targeted identification attacks mean that a huge dataset available is explicitly scanned for correlations that lead to a unique fingerprint of a single individual. Probable and provable results

refer the validity of the results gathered in big data. Economic effects of the big data paradigm are direct results of the exchange of datasets among business partners in advance.

In the general architecture of big data analytics, both distributed big data storing and parallel big data processing are driven by the big data 3V challenges. In addition to the 3V challenges, big data also faces new security and privacy challenges. If big data is not authentic, newly mined knowledge becomes useless. Recently, a new dimension, veracity, has been advocated to address the security challenges in big data. However, the study of privacy in big data is still in its early stage. Therefore, we focus ourselves on big data privacy and identify the privacy requirements of big data analytics as follows.

While big data creates enormous values for economic growth and technical innovation, we are already aware that the deluge of data also raises new privacy concerns. Thus, privacy requirements in big data architecture should be identified as deeply as possible to balance the benefits of big data and individual privacy preservation.

Privacy requirements in various stages of big data collection, storage and processing are:

Privacy requirements in big data collection: As big data collection takes place pervasively, eavesdropping is possible, and the data could be incidentally leaked. Therefore, if the collected data is personal and sensitive, we must resort to physical protection methods as well as information security techniques to ensure data privacy before it is securely stored.

Privacy requirements in big data storage:Compared to eavesdropping an individual's data during the big data collection phase, compromising a big data storage system is more harmful. It can disclose more individual personal information once it is successful. Therefore, we need to ensure the confidentiality of stored data in both physical and cyber ways.

Privacy requirements in big data processing:The key component of big data analytics is big data processing, as it indeed mines new knowledge for economic growth and technical innovation. Because big data processing efficiency is an important measure for the success of big data, the privacy requirements of big data processing become more challenging. We never sacrifice big efficiency for big privacy, and should not only protect individual privacy but also ensure efficiency at the same time. In addition, since inter big data processing runs over multiple organizations' data, big data sharing is essential, and ensuring privacy in big data sharing becomes one of the most challenging issues in big data processing. Therefore, it is desirable to design efficient and privacy-preserving algorithms for big data sharing and processing.

In recent years, we have witnessed plenty of privacy preserving techniques. privacy requirements in traditional analytics, they are not sufficient to satisfy the privacy requirements in big data analytics scenarios.

6 BIG DATASECURITY TECHNIQUES

Organizations used various methods of deidentification to ensure security and privacy. The most common solution to ensure security and privacy may be oral and written pledges. However, history has shown that this method is flawed. Passwords, controlled access, and twofactor authentication is low-level, but routinely used, technical solution to enforce security and privacy when sharing and aggregating data across dynamic, distributed data systems. Access permissions such as these can potentially be broken by both the intentional sharing of permissions and the continuation of permissions after they are no longer required or permitted.

More advanced technological solution is cryptography. The famous encryption schemes have AES and RSA. Recent revelations show that the National Security Administration (NSA) may have already found ways to break or circumvent existing Internet encryption schemes. Virtual barriers such as firewalls, secure sockets layer and transport layer security are designed to restrict access to data. Each of these technologies can be broken, however, and thus need to be constantly monitored, with fixes applied as needed.

Tracking, monitoring or auditing software is developed to provide a history of data flow and network access by an individual user in order to ensure compliance with security related. The limitation of this technology is that it is difficult and costly to implement on a large scale or with distributed data systems and users because it requires dedicated staff to read and interpret the findings, and the software can be exploited to monitor individual behavior rather than protecting data.

Thus the traditional de-identification techniques are not applicable in the era of Big Data since the de-identification technique widespread uses. The tasks of ensuring Big Data security and privacy become more difficult as information is increased. Computer scientists have repeatedly shown that even anonymized data can often be reidentified and attributed to specific individuals"

7 EXISTING PRIVACY PRESERVING TECHNIQUES

Given below are some existing privacy-preserving techniques, including privacy-preserving aggregation, operations over encrypted data, and de-identification techniques.

7.1 Privacy preserving aggregation

Privacy-preserving aggregation, which is built on some homomorphic encryption [5], is a popular data collecting technique for event statistics. Given a homomorphic public key encryption algorithm E(.), different sources can use the same public key to encrypt their individual data m_1, m_2, …, m_n into ciphertexts $c_1 = E(m_1)$, $c_2 = E(m_2)$, …, $c_n = E(m_n)$. By taking the sum aggregation as an example, these ciphertexts can be aggregated as $C = \prod_i^n = {}_1 c_i = E(\sum_i^n = {}_1 m_i)$. With the corresponding private key, the aggregated result $\sum_i^n = {}_1 m_i$ can be recovered from C. Obviously, privacy preserving aggregation can protect individual privacy in the phases of big data collecting and storing. However, since aggregation is purpose-specific, one-purpose aggregated data usually cannot be used for other purposes. Since its inflexibility prevents running complex data mining to exploit new knowledge, privacy-preserving aggregation is insufficient for big data analytics.

7.2 Operations over encrypted data

Currently, searching over encrypted data has been widely studied in cloud computing [6]. To keep sensitive documents private, documents and their associated keywords are encrypted and stored in a cloud server. When a user submits a "capability" encoding some query conditions, the server can return a set of encrypted documents that meet the underlying query conditions without knowing other details of the query. In such a way, the user can retrieve the desired data in a privacy-preserving way. Motivated bysearching over encrypted data, our first feeling is that we can also run operations over encrypted data to protect individual privacy in big data analytics. However, as operations over encrypted data are usually complex and time-consuming, while big data is high-volume and needs us to mine new knowledge in a reasonable time-frame, running operations over encrypted data is inefficient in big data analytics.

7.3 De-identification

De-identification is a traditional technique for privacy-preserving data mining, where in order to protect individual privacy, data should be first sanitized with generalization (replacing quasi-identifers with less specific but semantically consistent values) and suppression (not releasing some values at all) before the release for data mining. Compared to privacy-preserving aggregation and operations over encrypted data, deidentification can make data analytics and mining more effective and flexible [7]. However, many real examples indicate that data which may look anonymous is actually not after deidentification; for example, only (5-digit zip code, birth date, gender) can uniquely identify 80 percent of the population in the United States. Therefore, to mitigate the threats from reidentification, the concepts of k-anonymity, l-diversity, and tcloseness have been introduced to enhance traditional privacy-preserving data mining. Obviously, de-identification is a crucial tool in privacy protection, and can be migrated to privacy-preserving big data analytics. However, as an attacker can possibly get more external information assistance for de-identification in the big data era, we have to be aware that big data can also increase the risk of re-identification. As a result, de-identification is not sufficient for protecting big data privacy.

8 CONCLUSION

From the above discussion, we can see that:

1. Privacy-preserving big data analytics is still challenging due to either the issues of flexibility and efficiency or reidentification risks

2. However, compared with privacy-preserving aggregation and operations over encrypted data, de-identification is more fsible for privacy-preserving big data analytics if we can develop efficient and privacy-preserving algorithms to help mitigate the risk of re-identification.

With these two points in mind, future research work on big data privacy should be directed toward efficient and privacypreserving computing algorithms in the big data, and thesealgorithms should be efficientlly and output correct results while hiding raw individual data. In such a way, they can reduce the re-identification risk in big data analytics and mining.

REFERENCES

[1] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[2] Priya P. Sharma, Chandrakant P. Navdeti, (2014), " Securing Big Data Hadoop: A review of Security Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131

[3] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and itsApplications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.

[4]] K, Chitharanjan, And Kala Karun A. "A Review On Hadoop — Hdfs Infrastructure Extensions.".JejuIsland: 2013, Pp. 132-137, 11-12 Apr. 2013.

[5] P. Paillier, "Public-Key Cryptosystems based on Composite Degree Residuosity Classes," EUROCRYPT, 1999, pp. 223–38.

[6] M. Li et al., "Toward Privacy-Assured and Searchable Cloud Data StorageServices," IEEE Network, vol. 27, no. 4, 2013, pp. 1–10.

[7] A. Cavoukian and J. Jonas, "Privacy by Design in the Age of Big Data,"Office of the Information and Privacy Commissioner, 2012.K,

[8] Kilzer, Ann, Emmett Witchel, Indrajit Roy, VitalyShmatikov, and SrinathT.V.Setty. "Airavat:Security and Privacy for MapReduceM. Li et al., "Toward Privacy-Assured and Searchable Cloud Data Storage-Services," IEEE Network, vol. 27, no. 4, 2013, pp. 1–10.

[9] Chanchal Yadav, Shullang Wang, Manoj Kumar, (2013) "Algorithm and Approaches to handle large Data- A Survey", IJCSN, 2(3), ISSN:2277-5420(online), pp2277-54r, Ann, Emmett Witchel, Indrajit Roy, VitalyShmatikov, and SrinathT.V.Setty. "Airavat:Security and Privacy for MapReduce

[10] A. Cavoukian and J. Jonas, "Privacy by Design in the Age of Big Data," Office of the Information and Privacy Commissioner, 2012.

[11] R. Lu, X. Lin, and X. Shen, "SPOC: A Secureand Privacy-Preserving Opportunistic Computing Framework for Mobile-Healthcare Emergency," IEEE Trans. Parallel Distrib.Sys. , vol. 24, no. 3, 2013, pp. 614–24.

[12] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao, "Toward Efficient and Privacy-Preserving Computing in Big Data Era" IEEE Network July/August 2014