International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 2, August 2014)

Big Data Security Challenges: Dealing with too Many Issues

Rashmi N[1], Uma K M[2], Jayalakshmi K[3], Vinodkumar K P[4]

[1,2,3,4]*Assistant Professor, Dr. Ambedkar Institute of Technology, Bangalore, India*

Abstract— **Big data is magnified by high volume, high velocity, and high variety information assets. The scale of data and applications grow exponentially, and bring huge challenges of dynamic data monitoring and security protection. Big Data breaches will be big too, with the potential for even more serious reputational damage and legal repercussions than at present. Current technologies of privacy protection are mainly based on static data set, while data is always dynamically changed, including data pattern, variation of attribute and addition of new data. However, securing these large-scale data sets is typically beyond the reach of small businesses and it is increasingly posing challenges even for large companies and institutes. Thus, it is a challenge to implement effective privacy protection in this complex circumstance. Big Data expands the boundaries of existing information security responsibilities and introduces significant new risks and challenges. This paper introduces big data security issues and new Challenges. We discuss how storage and processing big data are affected by security and privacy factor. We also discuss few challenges like Secure computations in distributed programming frameworks, Secure Data Storage and Transactions Logs, Real-time security monitoring, Scalable and compassable privacy-preserving data mining and analytics and Cryptographically enforced access control and secure communication.**

Keywords- **Big Data, Security, Distributed Programming Frameworks, Real-time Security.**

I. INTRODUCTION

The term big data refers to the massive amounts of digital information companies and governments collect about us and our surroundings. Every day, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone. Big data is 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 infrastructures, with a diversity of software platforms, spread across large networks [3] of computers, also increases the attack surface of the entire system.

Traditional security [2] mechanisms, which are tailored to securing small-scale static (as opposed to streaming) data, are inadequate. For example, analytics [4] for anomaly detection would generate too many outliers. Similarly, it is not clear how to retrofit provenance in existing cloud infrastructures. Streaming data demands ultra-fast response times from security and privacy solutions. Big Data breaches will be big too, with the potential for even more serious reputational damage and legal repercussions than at present. A growing number of companies are using the technology to store and analyze petabytes of data including web logs, click stream data and social media content to gain better insights about their customers and their business. Security threats are becoming more aggressive and voracious. Governments and industry bodies are getting more prescriptive around compliance. Combined with exponentially more complex IT environments, security [2] management is increasingly challenging. Moreover, new "Big Data" technologies purport bringing advanced analytic techniques like predictive analysis and advanced statistical.

Techniques close to the security professional. Given the state of today's security systems, most organizations are a long way from using these types of advanced technologies for security management. Security Professionals need to get more value from the data already collected and analyzed. They also need a better understanding of both current issues and impending challenges related to data. Starting with a foundational set of data management and analytic capabilities enables organizations to effectively build and scale security management as the enterprise evolves to meet Big Data challenges [2]. However, securing these large-scale data sets is typically beyond the reach of small businesses and it is increasingly posing challenges even for large companies and institutes.

The goal of this paper is to provide big data security issues and new Challenges. We discuss how storage and processing [1] big data are affected by security and privacy factor.

We also discuss few challenges like Secure computations in distributed programming frameworks, Secure Data Storage and Transactions Logs , Real-time security monitoring, Scalable and compassable privacy-preserving data mining and analytics, Cryptographically enforced access control , secure communication, Granular access control, Granular audits Data provenance.

II. Security Issues And Challenges In Big Data

2.1 Secure Computations in Distributed Programming Frameworks

Distributed programming frameworks utilize parallelism in computation and storage to process massive amounts of data. A popular example is the MapReduce[5] framework, which splits an input file into multiple chunks. In the first phase of MapReduce, a Mapper for each chunk reads the data, performs some computation, and outputs a list of key/value pairs. In the next phase, a Reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper.

Use Cases

Untrusted mappers could return wrong results, which will in turn generate incorrect aggregate results. With large data sets, it is next to impossible to identify, resulting in significant damage, especially for scientific and financial computations. Retailer consumer data is often analyzed by marketing agencies for targeted advertising or customer-segmenting. These tasks involve highly parallel computations over large data sets, and are particularly suited for MapReduce[6] frameworks such as Hadoop. However, the data mappers may contain intentional or unintentional leakages. For example, a mapper may emit a very unique value by analyzing a private record, undermining users' privacy.

2.2 Secure Data Storage and Transactions Logs

Data and transaction logs are stored in multi-tiered storage media. Manually moving data between tiers gives the IT manager direct control over exactly what data is moved and when. However, as the size of data set has been, and continues to be, growing exponentially, scalability and availability have necessitated auto-tiering for big data storage management. Auto-tiering solutions do not keep track of where the data is stored, which poses new challenges to secure data storage. New mechanisms are imperative to thwart unauthorized access and maintain the 24/7 availability.

Use Cases

A manufacturer wants to integrate data from different divisions. Some of this data is rarely retrieved, while some divisions constantly utilize the same data pools. An auto-tier storage system will save the manufacturer money by pulling the rarely utilized data to a lower tier. However, this data may consist in R&D results, not popular but containing critical information. As lower-tier often provides decreased security, the company should study carefully tiering strategies.

2.3 Real-time Security/Compliance Monitoring

Real-time security[2] monitoring has always been a challenge, given the number of alerts generated by (security) devices. These alerts (correlated or not) lead to many false positives, which are mostly ignored or simply "clicked away," as humans cannot cope with the shear amount. This problem might even increase with big data, given the volume and velocity of data streams. However, big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing[1] and analytics of different types of data. Which in its turn can be used to provide, for instance, real-time anomaly detection based on scalable security analytics[4].

Use Cases

Most industries and government (agencies) will benefit from real-time security analytics, although the use cases may differ. There are use cases which are common, like, "Who is accessing which data from which resource at what time"; "Are we under attack?" or "Do we have a breach of compliance standard C because of action A?" These are not really new, but the difference is that we have more data at our disposal to make faster and better decisions in that regard. However, new use cases can be defined or we can redefine existing use cases in lieu of big data. For example, the health industry largely benefits from big data technologies, potentially saving billions to the tax-payer, becoming more accurate with the payment of claims and reducing the fraud related to claims. However, at the same time, the records stored may be extremely sensitive and have to be compliant with HIPAA or regional/local regulations, which call for careful protection of that same data.

Detecting in real-time the anomalous retrieval of personal information, intentional or unintentional, allows the health care provider to timely repair the damage created and to prevent further misuse.

2.4 Scalable and Composable Privacy-Preserving Data Mining and Analytics

Big data can be seen as a troubling manifestation of Big Brother by potentially enabling invasions of privacy, invasive marketing, decreased civil freedoms, and increase state and corporate control. A recent analysis of how companies are leveraging data analytics [4] for marketing purposes identified an example of how a retailer was able to identify that a teenager was pregnant before her father knew. Similarly, anonymizing data for analytics is not enough to maintain user privacy. For example, AOL released anonymized search logs for academic purposes, but users were easily identified by their searchers. Netflix faced a similar problem when users of their anonymized data set were identified by correlating their Netflix movie scores with IMDB scores. Therefore, it is important to establish guidelines and recommendations for preventing inadvertent privacy disclosures.

Use Cases

User data collected by companies and government agencies are constantly mined and analyzed by inside analysts and also potentially outside contractors or business partners. A malicious insider or untrusted partner can abuse these datasets and extract private information from customers. Similarly, intelligence agencies require the collection of vast amounts of data. The data sources are numerous and may include chat-rooms, personal blogs and network routers. Most collected data is, however, innocent in nature, need not be retained, and anonymity preserved. Robust and scalable privacy preserving mining algorithms will increase the chances of collecting relevant information to increase user safety.

2.5 Cryptographically Enforced Access Control and Secure Communication

To ensure that the most sensitive private data is end-to-end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. Specific research in this area such as attribute-based encryption (ABE) has to be made richer, more efficient, and scalable. To ensure authentication, agreement and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented.

Use Cases

Sensitive data is routinely stored unencrypted in the cloud. The main problem to encrypt data, especially large data sets, is the all-or-nothing retrieval policy of encrypted data, disallowing users to easily perform fine grained actions such as sharing records or searches. ABE alleviates this problem by utilizing a public key cryptosystem where attributes related to the data encrypted serve to unlock the keys. On the other hand, we have unencrypted less sensitive data as well, such as data useful for analytics. Such data has to be communicated in a secure and agreed-upon way using a cryptographically secure communication framework.

2.6 Granular Access Control

The security [2] property that matters from the perspective of access control is secrecy—preventing access to data by people that should not have access. The problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security. Granular access control gives data managers a scalpel instead of a sword to share data as much as possible without compromising secrecy.

Use Cases

Big data analysis and cloud computing[2] are increasingly focused on handling diverse data sets, both in terms of variety of schemas and variety of security requirements. Legal and policy restrictions on data come from numerous sources. The Sarbanes-Oxley Act levees requirements to protect corporate financial information, and the Health Insurance Portability and Accountability Act includes numerous restrictions on sharing personal health records. Executive Order 13526 outlines an elaborate system of protecting national security information. Privacy policies, sharing agreements, and corporate policy also impose requirements on data handling. Managing this plethora of restrictions has so far resulted in increased costs for developing applications and a walled garden approach in which few people can participate in the analysis. Granular access control is necessary for analytical systems to adapt to this increasingly complex security environment.

2.7 Granular Audits

With real-time security[2] monitoring, we try to be notified at the moment an attack takes place. In reality, this will not always be the case (e.g., new attacks, missed true positives).

In order to get to the bottom of a missed attack, we need audit information. This is not only relevant because we want to understand what happened and what went wrong, but also because compliance, regulation and forensics reasons. In that regard, auditing is not something new, but the scope and granularity might be different. For example, we have to deal with more data objects, which probably are (but not necessarily) distributed.

Use Cases

Compliance requirements (e.g., HIPAA, PCI, Sarbanes-Oxley) require financial firms to provide granular auditing records. Additionally, the loss of records containing private information is estimated at $200/record. Legal action – depending on the geographic region – might follow in case of a data breach. Key personnel at financial institutions require access to large data sets containing PI, such as SSN. Marketing firms want access, for instance, to personal social media information to optimize their customer-centric approach regarding online ads.

2.8 Data Provenance

Provenance metadata will grow in complexity due to large provenance graphs generated from provenance-enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive.

Use Cases

Several key security applications require the history of a digital record – such as details about its creation. Examples include detecting insider trading for financial companies or to determine the accuracy of the data source for research investigations. These security assessments are time sensitive in nature, and require fast algorithms to handle the provenance metadata containing this information. In addition, data provenance complements audit logs for compliance requirements, such as PCI or Sarbanes-Oxley.

III. Conclusion

Big data is not new concept but very challenging. Effectively managing and prioritizing the volume, variety, and velocity of data requires human insight, a multi-pronged approach, and multiple layers of defense.

In this paper, we discussed big data security issues and Challenges and also how storage and processing[1] big data are affected by security and privacy factor. We respectively discussed the key issues including multiple infrastructure tires for processing big data, non-scalability of real-time monitoring techniques that might be practical for smaller volumes of data, the heterogeneity of devices that produce the data, and policy restrictions that leads to ad hoc approaches for ensuring security and privacy .Many of the types in the list also serve to clarify specific aspects of the attack surface of the entire big data processing infrastructure that should be analyzed for these types of threats.

REFERENCES

[1] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, "Big data processing in cloud computing environment" in International Symposium on Pervasive Systems, Algorithms and Networks 2012.

[2] Disha H. Parekh, Dr. R. Sridaran ,"An Analysis of Security Challenges in Cloud Computing" in International Journal of Advanced Computer Science and Applications, Vol. 4, No.1, 2013.

[3] Jeff Whitworth , Shan Suthaharan ," Security Problems and Challenges in a Machine learning-based Hybrid Big Data Processing Network systems".

[4] Big Data Working Group ," Big Data Analytics for Security Intelligence In CLOUD SECURITY ALLIANCE" September 2013

[5] Dr.Siddaraju, Rahul M, Sowmya C L, Rashmi K, "Efficient Analysis of Big Data using MapReduce Framework", IJRDET, June 2014.

[6] Jens Dittrich JorgeArnulfo Quian´eRuiz ," Efficient Big Data Processing in Hadoop MapReduce".

[7] Jianqing Fan, Fang Han and Han Liu," Challenges of Big Data analysis",in National Science Review, 2014.

[8] The Right Tools for Smart Protection A Trend Micro White Paper, "Addressing Big Data Security Challenges", September 2012

[9] Advanced Technology & Engineering Research (IJATER) National Conference on Emerging Trends in Technology (NCET-Tech) ISSN No: 2250-3536 Volume 2, Issue 4, July 2012 1

[10] Ashalatha R¹ Faculty of Computer Science, Dayananda Sagar College of Engineering, Bangalore, "A SURVEY ON SECURITY AS A CHALLENGE IN CLOUD COMPUTING". E-mail: ashalatha.dsce@gmail.com

[11] Mesra, Ranchi," Big Data Analysis: Challenges and Solutions in International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV 269