International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(3), pp. 313-318
DOI: http://dx.doi.org/10.21172/1.73.543
e-ISSN:2278-621X

Security And Privacy Challenges In Big Data

Khantil Choksi[1], Niriksha Dalal[2], Mr. Kshitij Gupte[3] and Dr. Anjali Jivani[4]

Abstract- Nowadays, the obtainability of Big Data holds much promise to utilize the power of copious data sets and convert that power into transformations and advances in science, medicine, health care, education, and economic growth. While ensuring data security and privacy, challenges still remain of appropriate use of these massive data sets. These challenges are like vulnerabilities in public databases, protection against security breaches and data leakage etc. While managing large-scale, distributed data sets, the security and privacy policies throws a major challenge in tracking and monitoring data access and use in a dynamic, decentralized environment. Hence, the purpose of this paper is to elucidate the Big Data security and privacy challenges.

Keywords- Big Data, Security challenges, Privacy

I. INTRODUCTION

Big data is referred as the large-scale information management and analysis that exceed the capability of traditional data processing technologies. It is the massive amounts of digital information which companies and governments collect about human beings and our surroundings.

Big Data is differentiated from the traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety). The security and privacy issues are magnified by the above mentioned three ways. [1]

Former security mechanisms, which are mainly tailored for securing small-scale static data as opposed to streaming data, are insufficient. For example, analytics for anomaly detection would generate too many outliers. Likewise, it is not clear how to modify provenance in existing cloud infrastructures. Streaming data demands ultra-fast response times from security and privacy solutions.

In today's period, Big Data is used by industries of all levels which have access to Big Data and the means to employ it. Software infrastructures such as Hadoop enable developers to distribute storage and distribute processing of very large data sets on computer clusters, which easily leverage millions of computing nodes to perform data-parallel computing. [1] With the combination of the ability to buy computing power on-demand from public cloud providers, such developments greatly aggrandize the adoption of Big Data mining methodologies. Therefore, new security challenges have turned out from the coupling of Big Data with public cloud environments categorized by heterogeneous compositions of hardware with operating systems, and software infrastructures for storing and computing on data.

[1] *Department of Computer Science & Engineering, Faculty of Technology & Engineering, The M. S. University of Baroda, Vadodara, India*
[2] *Department of Computer Science & Engineering, Faculty of Technology & Engineering, The M. S. University of Baroda, Vadodara, India*
[3] *Department of Computer Science & Engineering, Faculty of Technology & Engineering, The M. S. University of Baroda, Vadodara, India*
[4] *Department of Computer Science & Engineering, Faculty of Technology & Engineering, The M. S. University of Baroda, Vadodara, India*

II. BIG DATA SECURITY AND PRIVACY CHALLENGES

According to recent article by Cloud Security Alliance (CSA) [1], there are mainly ten challenges in the field of Big Data security and privacy as mentioned below:

1) Secure computations in distributed programming frameworks
2) Security best practices for non-relational data stores
3) End-point input validation and filtering
4) Real-time security monitoring
5) Privacy-preserving data mining and analytics
6) Cryptographically enforced data centric security
7) Granular access control
8) Secure data storage and transactions logs
9) Granular audits
10) Data provenance

In the Big Data system, challenges can be categorized into four aspects namely: [1]-[2] [Fig-1]

- Infrastructure Security
- Integrity and Reactive Security
- Data Privacy
- Data Management

Fig. 1- Categorized Challenges

1) Secure Computations in Distributed Programming Frameworks :

Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Distributed programming frameworks harness parallelism in computation and storage to process massive amounts of data.
For instance, MapReduce [Fig-2] is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. It usually divides the input data sets into independent lumps, which are processed by map tasks in a complete parallel manner. In the first phase of MapReduce, a Mapper for each lump reads the data, performs some calculation and outputs a list of key/value pairs. In the later phase, a Reducer combines the values belonging to each unique key and outputs the result.

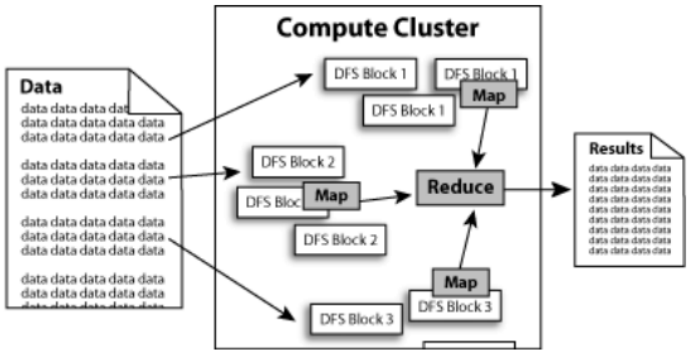

Fig. 2- MapReduce Architecture

There are two main attack prevention measures [3] : securing the mappers and securing the data in the presence of an untrusted mapper. Untrusted mappers could return anonymous results, which will in turn generate incorrect cumulative results. In scientific and financial computations, using large-scale data sets, it is implausible to identify results of significant damage. For targeted advertising or customer-segmenting, retailer-consumer data is often analyzed by marketing agencies.[1] These tasks involve high amount of parallel calculations over large data sets, and are particularly suited for MapReduce frameworks such as Hadoop. Although, the data mappers may contain intentional or unintentional leakages. For instance, a mapper may leak a very distinct value by analyzing private data, undermining users' privacy.

2) Security Best Practices for Non-Relational Data Stores :

An issue with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform many forms of non-SQL processing, like data mining and statistical analyses. Non-relational data stores known as NoSQL databases are still struggling with respect to security infrastructure. For example, well-conditioned solutions to NoSQL injection are still not mature. Each NoSQL DBs were constructed to handle various challenges posed by the analytics world and hence security was never part of the model at any point of its design stage. Most developers using NoSQL databases generally integrate security in the middleware.[4] No support is provided by NoSQL databases for enforcing it explicitly in the database. Such security practices poses additional challenges. In terms of accommodating and processing huge volume of data, organizations dealing with big unstructured data sets may gain advantage by migrating from a traditional relational database to a NoSQL database. Hence, the security of NoSQL databases relies in external enforcing mechanisms.
In order to lower down the security incidents, the company must go through security policies for the middleware adding items to its engine and toughen NoSQL database itself to match RDBs without compromising on its operational features.

3) End-Point Input Validation and Filtering :

Many big data collects data from variety of sources, such as end-point devices. There are two fundamental risks involved in data collection process. First, when we collect data from millions of hardware devices and software applications in an enterprise network, a major challenge in the data collection process is input validation. Validation of the input data sources poses the questions such as whether the input data is malicious or not and furthermore, filtering the malicious input from the collection. Second, filtering of data is also one of the major concern since amount of data collection in Big Data makes it implausible to validate and filter data on the fly. Former signature based data filtering may fail to solve the input validation and data filtering problem completely.

4) Real-time Security and Compliance Monitoring :

Real-time security monitoring is always challenging, given the number of alerts generated by security devices. These alerts lead to many false positives, which are mostly ignored as humans cannot cope with the shear amount. Given the volume and velocity of data streams, this problem might even increase with big data. Nevertheless, big data technologies might also provide an opportunity that is these technologies do allow for fast processing and analytics of different types of data and this in turn can be used to provide real-time anomaly detection related to scalable security analytics. Most industries will benefit from real-time security analytics, although the use cases may differ. For instance, the health industry largely profits from big data technologies, potentially saving billions to the tax-payer, becoming more accurate with the payment of claims and reducing the fraud. In this regard, a key challenge is that by whom it is being accessed and which resource is being accessed at what time.

5) Privacy-Preserving Data Mining and Analytics :

Big data can be seen as enabling invasions of privacy, invasive marketing, decreased civil freedoms, and increase state and corporate control. Anonymizing data for analytics is not enough to maintain user privacy. For example, Netflix (American multinational entertainment company which provides streaming media and video on demand online) faced a problem when users of their anonymised data set were identified by correlating their Netflix movie scores with IMDB scores [1]. Hence, it is important to establish guidelines and recommendations for preventing inadvertent privacy disclosures.

User data mustered by companies and government agencies are constantly mined and analyzed by inside analysts and also by outside contractors. A malicious insider or unauthorized partner can abuse these datasets and fetch private information from customers. Likewise, intelligence agencies require the collection of vast amounts of data. Robust and scalable privacy preserving algorithms will increase the probability of collecting relevant information to accentuate user privacy.

6) Cryptographically Enforced Access Control and Secure Communication :

To make sure that the most sensitive private data is fully secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. To ensure authentication, agreement and fairness among the distributed entities, a communication framework [2] which is cryptographically secured has to be implemented.

Sensitive data is generally stored unencrypted in the cloud. The main problem to encrypt data is the all-or-nothing retrieval policy of encrypted data, which restrain users from easily performing fine grained actions such as sharing records or searches. Attribute-based encryption (ABE) [1] alleviates this problem by utilizing a public key cryptosystem where attributes related to the data encrypted serve to decrypt the keys. On the other hand, the unencrypted less sensitive data useful for analytics, has to be communicated in a secure and agreed-upon way using a cryptographically secure communication framework.

7) Granular Access Control :

Present solutions of Big Data are designed for performance and scalability, keeping almost no security in mind. Traditional relational databases have good security features in terms of access control, users, tables and rows and even at cell level. Still, various fundamental challenges prevent Big Data solutions to provide comprehensive access control. The major problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive group to guarantee security. [1] Granular access control is necessary for analytical systems to adapt to this increasingly complex security environment.

Keeping log of roles and authorities of users is one of the threat along with maintaining access labels across analytical transformations. Big data analysis and cloud computing are increasingly focused on handling diverse data sets, both in terms of variety of schemas and requirements. Legal and policy restrictions on data come from numerous sources. [1] Privacy policies, sharing agreements, and corporate policy also impose requirements on data handling. Managing this excess 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.

8) Granular Audits :

When it comes to real time security monitoring; notification of an attack at the moment it takes place is the goal. In real life, this is not always the case. In other to determine missed attack, auditing of information is necessary. Information from auditing is very key to understanding what occurred and what went wrong, for compliance, regulations and forensic investigation. Auditing is not something new, but big data tend to extend its reach because of its volume and sometimes distributed processes. Auditing capabilities need to be implemented across big data infrastructure depending on the auditing features enabled for the infrastructure components. Examples include syslog on routers [1], application logging and enabling logging on the operating system.

9) Secure Data Storage and Transactions Logs :

The security problems on secure data storage and transaction logs is related with data storage, management, and processing. Security not available in big data storage may corrupt the data due to unauthorized access. While transferring the data from one source to another, unauthorized access may cause tampered data to get delivered to end users. Data and transaction logs are stored in multi-tiered storage media. Moving the data manually between tiers, gives the developer direct control over preciously what data is moved and when. However, as the data set size grows exponentially, scalability and availability have necessitated auto-tiering for big data storage management. [5] Auto-tiering solutions have no track of where the data is stored, which questions new challenges to secure data storage.

For instance, an organization wants to integrate data from different divisions. Much of this data is hardly retrieved and accessed, while some divisions constantly utilize the same data pools. An auto-tier storage system will save the organization money by pulling the hardly utilized data to a lower tier. But, this data may contain not popular, but critical information. As lower-tier often provides decreased security, the company should review tiering strategies meticulously.

10) Data Provenance :

In its strongest form, data provenance supports information and process integrity by documenting the entities, systems, and processes operating on and contributing to data of interest. This serves as an

unalterable historical record of the data's lifetime and its sources. Analysis of large provenance graphs to detect metadata dependencies for security and confidentiality applications is computationally intensive.

Various key security applications require the history of a digital record – such as details about its creation [1]. Instances include evaluating insider trading for financial companies or to determine the precision of the data source for research investigations. These security assessments are critical in terms of time, and require fast algorithms to handle the provenance metadata containing this information.

III. CONCLUSION

Through proper analysis of both streaming and static large data sets, we can make better advances in many scientific and medical disciplines and profitability for many enterprises. It is pragmatically implausible to imagine the next application without it consuming data as well as creating new forms of data, and containing data-driven algorithms. The challenges introduced by security, access control, compression, encryption and compliance must be addressed in a systematic way as computing environments become cheaper, application environments become networked, and system and analytics environments become shared over the cloud. This paper has highlighted the top security and privacy problems that need to be addressed if we are to make Big Data processing and computing infrastructure more secure. We believe that this paper will stimulate action in the research and development community to collaboratively focus on the barriers to higher security and privacy in Big Data platforms.

REFERENCES

[1] A Cloud Securty Alliance Collaborative research, "Expanded Top Ten Big Data Security and Privacy challenges" , April 2013.

[2] A community White paper developed by leading researchers across united states, "Challenges and Opportunity with Big Data", Feb. 2012.

[3] I. Roy, S. T. V. Setty, A. Kilzer, V. Shmatikov and E. Witchel, "Airavat: security and privacy for MapReduce" in USENIX conference on Networked systems design and implementation, pp 20-20, 2010.

[4] L. Okman, N. Gal-Oz, Y Gonen, E. Gudes and J Abramov, "Security Issues in NoSQL Databases" in TrustCom IEEE Conference on International Conference on Trust, Security and Privacy in Computing and Communications, pp 541-547, 2011.

[5] Mr. Yannam Apparao, Mrs. Kadiyala Laxminarayanamma, "Security Issue on Secure Data Storage and Transaction Logs In Big Data" in International Journal of Innovative Research in Computer Science & Technology (IJIRCST), May 2015