International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(3), pp.235-239
DOI: http://dx.doi.org/10.21172/1.83.035
e-ISSN:2278-621X

CHALLENGES OF BIG DATA PRIVACY AND SECURITY: A REVIEW

Dr.Shahnaz Fatima[1], Dr. Ranjana Rajnish[2] and Dr. Parul Verma[3]

Abstract- Big data is all about voluminous data which is flowing from various sources. Now day's data is streaming from various channels like social media, electronic devices, sensors and many others. Handling such huge amount of data is getting difficult for the data users. It is not defined anywhere as such that how much volume of data will be considered as "big" data, but handling of such data requires lot of new tools and techniques to process it. Big Data is become a business issue. The business demands data and information for strategically analyzing the current business demands and market trends. Maintaining the privacy and security of Big Data is a very critical issue. The volume and variety of big data alleviates the standard of security required for it. The currently used security mechanisms like firewalls and DMZs are not sufficient for Big Data as the data is coming across the limited range like within an organization. The basic motive of maintaining security is to preserve its confidentiality and integrity. This paper gives an overview of Software Defined Networking(SDN).This paper also highlights big data -specific security and privacy challenges and its solutions.

Keywords – Big Data, SDN, Privacy, Security

I. INTRODUCTION

Most definitions of big data focus on the size of data in storage. Size matters, but there are other important attributes of big data, namely data variety and data velocity. The three Vs of big data (volume, variety, and velocity) constitute a comprehensive definition, and they bust the myth that big data is only about data volume.

Figure 1: The Three V's of Big Data

Big data deals with the variety of a data. Different variety of data needs different analytic logic to produce results which work as an insight for the business. Variety refers to the source of data now days the sources are diversified

[1] *Amity Institute Of Information Technology Amity University,Lucknow,UP,India*
[2] *Amity Institute Of Information Technology Amity University,Lucknow,UP,India*
[3] *Amity Institute Of Information Technology Amity University,Lucknow,UP,India*

like emails, photos, videos, monitoring devices, PDFs, audio, etc. Such diversified unstructured data creates lot of issues in storing it as well as analyzing it also.To harness the power of Big data one needs an infrastructure and technologies which can deal with huge volume and variety of data as well as can draw inferences from it. There are various technologies for "Big Data Analysis" given by various vendors. The technologies are growing rapidly with the growing market of Big Data Analytics.

II. Privacy and Security in Big Data

Security functions required for the Big Data should be capable enough to manage heterogeneous data coming from the diverse hardware, operating system and network domain. In such a complicated scenario where diversity and volume is the big issue Software Defined Networking (SDN) is found to be suitable for the efficient deployment of Big Data Secure services on the top of heterogeneous infrastructure.

A. Software Defined Networking (SDN) for Big Data

The volume of data is a major issue in Big Data infrastructure development. The varied sources like – social networking sites, public sensors, GPS information and many more are creating unprecedented amount of data. This data cannot be easily captured and processed by using traditional data management systems. SDN comes in picture, which is emerging network architecture adaptable by many technologies. The SDN has the potential to handle multi-domain, automated, guaranteed bandwidth service that big data needs. SDN is reforming the way we design and manage networks[1]. There are two major components of SDN –

Control Plane: It decides how to handle the traffic in the network[1]. It basically works as decision maker. The control plane have direct control over the various data-plane elements like routers, switches and other middleboxes with the help of a well-defined Application Pro-gramming Interface (API)[1]. OpenFlow is one of such API which specifies number of packet handling rules.

Data Plane : It forwards traffic according to decisions that the control plane makes[1]. It basically works as a packet forwarder. The data planes are programmable in nature and can be maintained, controlled and programmed from a central entity.

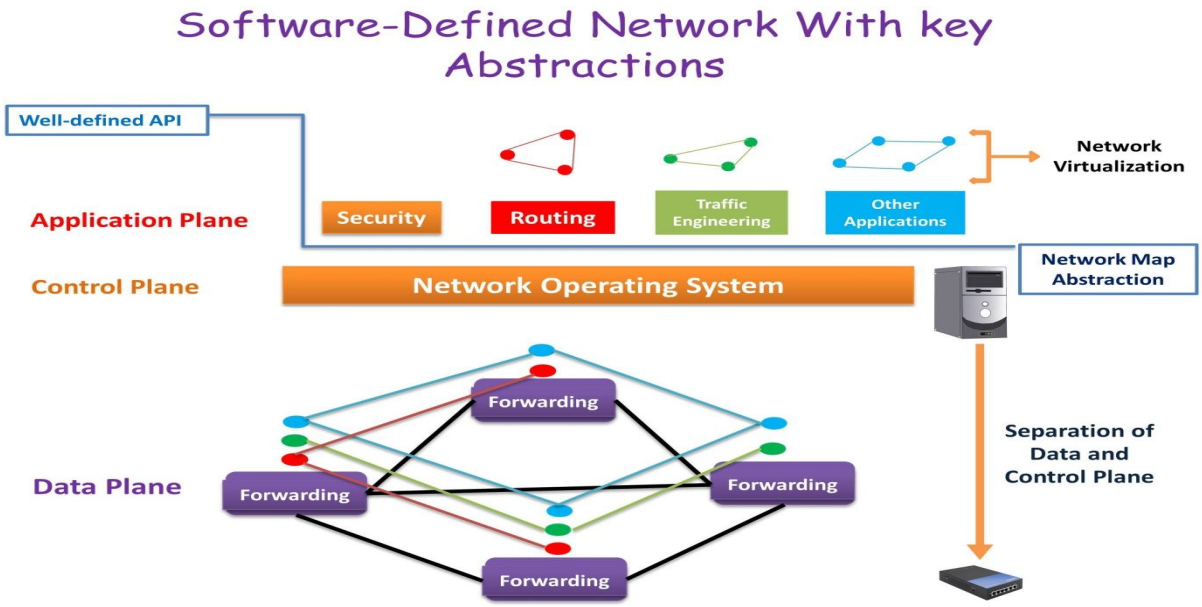

Figure 2: SDN Framework[4]

The basic concept of SDN is to separate the control plane from the data plane[2]. SDN can be easily managed centrally by the managers by configuring, programming, securing and optimizing the network services. It facilitates the network administrators to dynamically adjust the traffic flow in the network as per the changing requirements.

SDN is simply a concept of separating control plane from the data plane. OpenFlow is a communication interface between the two. It allows direct access and manipulation of the forwarding plane of devices such as switches and routers. It can be considered as a protocol which is used in switching devices and controllers interface.

The industry experts are trying to visualize the potential of SDN for big data technologies. Big data-processing software requires agile, multi-domain, centrally managed architecture. SDN addresses the stasis and lack of scalability of current networks by:[3]

- Allowing for changing traffic patterns
- Providing functionality to applications that access geographically distributed databases and servers through public and private clouds
- Providing access to bandwidth on demand

SDN is capable of meeting the demands of Big Data infrastructure as it keeps on growing continuously in size and complexity along with growing number of diverse users. Both the control plan setup along with data plane decisions can be determined algorithmically or with respect to rules determined by the features or statistical makeup of very large data sets[3].SDN is considered as the promising networking infrastructure for the Big Data technologies. There are many issues regarding the implementation and interaction with the big data is still unanswered.

B. Challenges of Privacy and Security

Big Data faces high level of challenges while dealing with the privacy and security of voluminous and versatile data. Traditional security mechanisms are not found adequate for the current scenario of Big Data. The challenges can be categorized as follows-

i. Security of Infrastructure

Security of Distributed Processing of Data: The massive amount of data is managed by distributed programming framework using parallelism. One of the examples of such processing is a "MapReduce" which splits the input file into smaller chunks. The mapper reads the chunks and after processing, produces list of key/value pairs. The reducer combines the values belonging to each distinct key and produces the result. The attack prone issues are- securing the mappers and securing the data from untrusted mappers. Untrusted mappers could result into incorrect aggregate results.

Security of Non-Relational Data Stores: NoSql databases are still not mature enough and their security is at stake as no such robust solutions for NoSql injection have been evolved. The evolution of NoSql database was to majorly beat the challenges of the analytics world. That time security was never the agenda for the NoSql database design. The design model of NoSql does not provide any support for security; the developers have to take care of this aspect and need to introduce security as a middleware.

ii. Maintaining Data Privacy

Preserving Data Privacy: Big data faces the biggest challenge in maintaining privacy of data . Users data is collected by private companies/ government agencies , which is further mined and analyzed by insider analyst, business partners or outside contractors. Any insider or malicious partner can breach there datasets and extracts crucial information of the customers.

Cryptographically Secure Communication: Big data access needs to ensure that the sensitive private data is highly secure and only accessible to the authorized parties. This needs that data should be encrypted based on some access policies. Sensitive data in cloud is stored in unencrypted form. The encrypted data will bound the users in performing fine grained processing. ABE (Attribute Based Encryption) provides solution for this problem by using public key crypto-system, where particular attributes serve as a key to unlock the data.

Granular Access Control:It is about preventing access of data by people that should not have access to it. There are numerous restrictions from various government/ private organizations, for access of data. The major issue with coarse grained processing is to sweep the data into more restrictive category which can be shared. The role of Granular Access is to provide control to data managers to share as much data without breaching security.

iii. Integrity and Data Management

Transaction logs and Secure Data Storage: Data and transaction logs are maintained in a multi-tiered storage media. The IT manager can manually move the data between the tiers, however with the increasing volume of data manually doing this task is quiet difficult. The autotiering facilitates the same job in big data storage management. Autotiering does not keep track of where the data is stored which poses a challenge. The data in a multi-tier system is categorized into a frequently retrieved data and rarely retrieved data. The auto-tier storage system keeps the rarely retrieved data to a lower tier. The lower tier is given less security hence the organizations should be careful in tiering strategies.

Granular Audits: The requirement of the real-time monitoring system is to report the attack as it takes place. However practically it is not possible all the time, in that case to get the bottom details of the attack we need audit information. Audit information is necessary to understand what happened and what went wrong.

Data Provenance: Several applications require the history of a digital record for ex- details about its creation, usage etc. Such security assessments are time sensitive, and require fast access to the metadata containing the information. The complexity of provenance data is growing due to large performance graphs generated by big data applications.

iv. Reactive Security

Validation: Bring your own deice (BYOD) model poses a challenge for input validation and filtering. In enterprise settings data is collected from millions of hardware devices and software applications in a network. Validating the input is a biggest challenge. Identifying and filtering the malicious data is a key challenge for the enterprise network.

Security level in Real Time: Real time security analytics will benefit many private and government agencies. Answers of such questions like “who is accessing data and at what time”; “Are we under attack”; “Do we breach the rule C because of action A”. Such queries are not new but with the volume and variety of data we need to make faster and better decisions in this regard. Real-time security is always been a challenge. The challenge is bigger with the increasing in size, velocity and variety of data.

III. SOLUTIONS FOR THE CHALLENGES OF BIG DATA SECURITY AND PRIVACY

There is no such single solution for the challenges discussed of Big Data security. However there can be cascading solutions to fight the diversified challenges of Big Data. The need is to understand the complex security issues to protect both structured and unstructured data. Following solutions are suggested to overcome such privacy and security issues-

1. The most basic solution is to encrypt the data wherever it is residing. However as the size grows the processing of such encrypted data gets cumbersome.
2. The heterogeneous data cannot be secured by using the common policy. The requirement is to have the data specific security policy and to implement the policy right over the place of origin of data to prevent data leakage.
3. Authentication process verifies user or system identity. Identity authentication should be full proof in order to maintain security and privacy of the data.
4. Access control implementation which specifies the access control privileges enhances the security of a system.

5. Storing the keys in the local disk drive put it in danger, as it can be easily collected by the platform administrator or an attacker. The secure way is to use key management service to distribute keys and certificates and manage different key for each group, application and user.
6. Log files should be audited on periodic basis in order to meet security requirements. Log files provide the detailed information about the authorized and unauthorized access of data.
7. Implement secure communication between nodes and applications. This requires a SSL (Secure Socket Layer)/ TLS (Transport layer security) implementation to establish secure communication between two ends.
8. The legal requirements of data handling should be implemented.

Maintaining the privacy of data in Big Data scenario is quiet a challenging task. The great feat is of unauthorized access of private data of the users floating from diversified channels. To protect privacy there are two common approaches – First to restrict data access to the limited group of end users by issuing access control certificates. Second approach is to make data anonymous so the sensitive data cannot be pinpointed.

IV.CONCLUSION

Big Data refers to volume, variety and velocity of data. Big data handles the data set which is so large and complex and it is impossible to handle such volume and complex data with ordinary software tools. The big data is also categorized into Operational and Analytical data. To harness such variety and volume of data various technologies are being introduced. This paper discussed privacy and security of Big Data that need to be addressed for making big data processing more secure.

REFERENCES

[1] Feamster, Nick, Jennifer Rexford, and Ellen Zegura. "The road to SDN : an intellectual history of programmable networks", ACM SIGCOMM Computer Communication Review, 2014.

[2] http://ir.lib.ncu.edu.tw/handle/987654321/65538

[3] http://www.nec.com/en/global/ad/insite/article/bigdata02.html

[4] Introduction to Software Defined Network (SDN) , Hengky “Hank” Susanto, Sing Lab, HKUST

[5] “SoftwareDefined Networking: A Comprehensive Survey”, D. Kreutz , F. Ramos, et el. 2015.

[6] Survey on Software Defined Networking”, W. Xia, Y. Wen, et el. 2015