SURVEY OF SECURITY ISSUES IN BIG DATA

Syeda Ambreen Zafar, Muhammad Khalid Khan
College of Computing & Information Sciences, PAF-KIET
syedaambreenzafar@gmail.com, khalid.khan@pafkiet.edu.pk

ABSTRACT *– Big data has captured the attention of the world in recent years as it offers new tools and techniques to handle massive datasets. Various tools and techniques are developed for big data but these tools lack in handling security issues. The security issues in big data are different from traditional database security issues as the nature of work is different in big data. Various traditional mechanisms such as encryption algorithms and authentication techniques are used with big data tools to address security issues but still there are many open issues in this area. The paper presents a survey of big data security issues and mechanisms.*

Keywords – Big Data, Security Issues, Hadoop

I. INTRODUCTION

Big data are large data sets that are scalable, random and of different types that traditional business intelligence tools or applications failed to handle. Many challenges have been faced these days in different sectors mainly in telecomm companies or in social networks like analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The big data infrastructure provides a model that facilitates to have the better concept of big data components [1].

The Big Data architecture is distributed on five levels: "Data sources", "Integration process", "Data storage", "Analytics and computing models" and "Presentation". The "Data sources" layer consists of distinct data sources, from sensor streaming data, to structured information such as relational databases, and to any type of random data. The "Integration process" layer is concerned with gaining data and integrating the datasets with the important data pre-processing operations. The "Data storage" layer consists of a bunch of resources such as distributed file systems, RDF stores, NoSQL and NewSQL databases that is appropriate for the persistent storage of a large number of datasets. The "Analytics and computing models" layer compresses various data tools, such as Map Reduce, which run over storage resources and include the data management and the programming model. The "Presentation" layer facilitates the visualization technologies [1].

II. BIG DATA SECURITY ISSUES

In this section, general susceptibilities are discussed that are present in big data framework. It also highlights security issues that are normally overlooked by the big data technologies. The security attacks are also discussed in this section.

A. Drawbacks & Security Encounters in Big Data

Due to such huge volumes of data, it is difficult to verify the source and credibility of data which increases the uncertainty level. Therefore, for performance improvement, the data is available on distributed levels which results in making sensitive information approachable to any illicit person on the network.

In such a dispersed environment where data comes with different natures and sources whether it is a real time data or of transactional nature or some social site data, it is hard to secure the data. Not only it has diverse nature but it also it has a certain velocity like streaming over the internet or voice and video calls. Different companies share information for applying analytics and target marketing purposes. Therefore, it is almost impossible to maintain privacy of a sensitive data [2].

These aforesaid flaws create many security challenges that are broadly divided in the following four categories [3]:

Network level: The challenges that can be divided under a network level deal with network protocols and network security, such as distributed nodes, distributed data, Inter-node communication.

Authentication level: The challenges that can be divided under user authentication level deals with encryption/decryption techniques, authentication methods such as authentication of applications, nodes, and logging.

Data level: The challenges that can be divided under data level deals with data integrity and availability such as data protection and distributed data.

Generic types: The challenges that can be regarded as under general level are traditional security tools, and use of different technologies.

Peculiarly, big data security challenges are given below [4,5]:

Secure Computations in Distributed Programming Frameworks: In a map and reduce framework, corrupt mappers could generate wrong or inaccurate output which produces aggregate wrong results.

Secure Storage and Transaction Logs: Data and transaction logs are stored in multi layered infrastructure. In big data environment, data is stored via an auto-tiering process in which it is difficult to find as to where the data is stored. Which produces new challenges to secure storage.

Real-time Security/Compliance Monitoring: Real-time security 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.

Scalable Privacy-Preserving Data Mining and Analytics: Through data mining and applying analytics on large datasets for the purpose of target or invasive marketing which allows invasion of privacy can be one of the biggest challenge that increases state control and decreases civil freedom.

Cryptographically Enforced Access Control and Secure Communication: Sensitive data is routinely stored unencrypted in the cloud. The main problem to encrypt data, especially large data sets, is the all-or-nothing recovery

policy of encrypted data, restrict users to perform fine grained actions, such as sharing records or searches.

B. Security Issues in Big Data

Most of the big data technologies segregate all jobs or tasks within the framework into many systems for quick processing. So a small amount of effort is given by a single system , but it means a lot more systems where security issues can crop up [6], such as:

Non-relational data stores: Think NoSQL databases, which by themselves usually lack security

Storage: In big data architecture, the data is stored in multiple layers as per their business requirements vs cost. Therefore, locking down storage means a tier conscious strategy.

Endpoints: Logging that is drawn from the endpoints, which legitimacy needs to be verified, otherwise analysis will not be credible.

Real-time security/compliance tools: These produce enormous amounts of data where a false positive must be identified so a focus can be maintained on true security violation.

Data mining solutions: In big data environments, patterns are sought that helps in various business strategies. Therefore, it needs to be secured not only from external attacks but also from potential employers who intends to abuse user network rights to gain sensitive information.

Access controls: It is necessary to provide an environment where user privileges and rights are clearly defined and properly validated in an encrypted form.

Security Attacks in Big Data

Big data environment is pruned to many security attacks [2,7,8]. Some of them are:

Attack on Database: Joining databases leads to the leakage of sensitive information as well.

Personal identification attack: when any database is connected with the other then, personal data becomes vulnerable. When search for any specific information then other personal information associated with it also gets exposed which leads to the privacy violation.

Advanced Malicious Methods: Malware created today often undergoes quality control procedures. Cybercriminals test it on numerous machines and operating systems to ensure it bypasses detection.

Denial of Service: Data or network applications are denied or become inaccessible for even legitimate users by overloading server or bombarding them with excessive number of queries.

Input Injection: This attack involves inserting malicious statements into Big Data components (e.g., Hive or MapReduce) which can give an attacker unrestricted access to an entire database.

Malware: Cybercriminals, state-sponsored hackers, and spies use advanced attacks that blend multiple tactics—such as spear phishing emails and malware.

III. BIG DATA SECURITY MECHANISMS

The section presents various security mechanisms that promise to secure a typical big data environment [2].

Node authentication: Kerberos authentication is helpful for verifying user to user or node to node communication and keep illicit nodes and applications out of the cluster. It is one of the best mechanism so far for hardening the infrastructure, also protects web console access. It is very useful for distributed environment like Hadoop framework.

Layered encryption in files: File encryption prevents the data from various threats in application security. File encryption protects its content, in case of data breach or if files are stolen. It provides stable protection across various platforms regardless of its type. It is also scalable and a nominal way to handle several security threats.

Key management: Distributing keys and certificates using akey management service for each application or user. It needs some extra configuration and commercial key management products for scalable network. Mostly encryption controls rely on certificate or key security.

Activity logs: To identify any suspicious *activity, unusual behavior or node failure or any kind of malfunctioning,* you need a record of activity. Here big data facilitates large volumes of log data.

Use of SSL/TLS: Between nodes or nodes and applications secure socket layer (SSL) or transport layer security (TLS) should be implemented. There are few service providers who offer secure communication otherwise user needs to integrate these services.

Keyword search method: Keyword search method enable customers to search encrypted protection data. New search patterns are used such as rank or subset search.

Combined cloud security: Hybrid cloud bridges the communication between public and private cloud. Its processing is done at the block level instead of pixel level, which facilitates the faster communication.

Dual control: The networking management team should be divided into two separate groups: Network Group and Security Group. They have their separate passwords or keys for access. Network group must not be allowed to access log server or must not be allowed to have passwords for any security related content.

Behavior Profiling: The legitimate users on the network have a unique pattern of using the network like the specific type of applications they use or specific files they have. So if some malicious user enters the network, he must have a different pattern than other legitimate users. Which gives a chance to identify any activity that deviates from the usual patterns.

Related Work

The nature of big data is different than traditional database due to its high volume, variety and velocity. The big data includes all different types of data which is growing exponentially. It brings big challenges of monitoring the data, its privacy protection and generating threats as the uncertainty increases with data. The big data is dynamic and it changes constantly including its patterns and attributes. Which makes it impossible for small scale companies to handle such data. Even for larger companies who use traditional business intelligence techniques big data is a huge challenge. Granular Access Control provides facility to share data in chunks which helps in maintaining secrecy within those components. With real-time security monitoring, attack

Table 1: Summary of Big Data Security Attacks, Challenges, Issues and Methods

Big Data Security			Big Data Security Attacks	Big Data Security Challenges	Big Data Security Methods
Architectural	Operational		Attack on Database	Secure computation in distributed prog framework	Authentication using Kerberos
	Big Data Technology		Human Identification Attacks	Secure Data Storage and transaction logs	Layer encryption for files
Distributed	Batch Processing	Stream Processing	Innovative Malicious Methods	Real time Security/Compliance Monitoring	Key Management usage
Redundant	Access Control	Netflow Monitoring	Input Injection	Scalable privacy preserving Data mining/Analytics	Create logs of activities
Elastic Data Repository	Encryption Techniques	Behaviour Profiling	Denial of Service	Cryptographically enforced Access control & Secure computation	Use fo SSL/TLS
	Separation of Duties		Malware	**Big Data Security Issues**	Type Based keyword search
				Non-Relational Data Stores	Security via Hybrid Clolud
				Storage	Dual Control
				End points	Beehive:Behaviour Profiling
				Real time Security/ Compliance tools	
				Data Mining Solutions	

can be identified the moment it occurs. In order to investigate the attack or any threat we need to perform audit trails which is a common old practice, but its scope in a distributed environment might be different than that of a traditional practice [4].

Missed attacks can be identified with the help of Granular auditing. It can also help to identify when and how those attacks occurred and what content has been compromised and how to prevent those attacks in the future. This enormous amount of data must be protected which can be used to address security issues. The primary concern of data provenance is Meta data that is details about data itself. Which can tell about the data source, its purpose who accessed it etc. This type of data must be monitored in a real time in the middle of the attack. This activity must be meticulously examined to ensure they don't become their security issue itself [6].

The threat landscape is evolving with the exponential increase in number of threats. The main reason is that cybercriminals have become more professional and number of sophisticated tools they use, which means the environment for software security companies has become more challenging at an exceptional level. Therefore, preventing the system from the assault of cyber threats is not an easy job. If the intrusion detection system or incident response methodologies are not appropriate, the consequences would be unexpected. To take effective measures, the most suitable strategy and the right combination of methodologies must be adopted. In order to do efficient processing, one must have an Expert understanding of threats. For that matter, companies must examine how data is organized, understanding and analyzing complex relationships, using specialized algorithms and customized models to manage and protect data [7].

The implementing security controls on a layered architecture while keeping the holistic view on the entire network using actionable intelligence to prevent any suspicious activity. Searchable encryption techniques not only increases the operability of the network but also facilitates the privacy of the data. This method enables customers to search from encrypted data. Authorized users are able to store and query the data without decrypting all the files. There is also a concept of hybrid cloud has been proposed where sensitive data is filtered out and kept in a trusted private cloud while the non-sensitive in a public cloud. But the problem arises when the data has large volume and continuously growing. The cloud setup on this level with the variety of APIs also makes the entire system vulnerable [9].

In an enterprise network, graph implication approach on a vast level is introduced because the behavior of a malicious user must vary from any legitimate user over the network. Because they form a specific pattern that an illicit user can deviate from. Which can be easily identified. To tackle the implications that users can produce, regulatory policies and certain mechanisms are needed [10].

Companies are applying big data analytics for business development which has become an effective way to gain the useful insight of any data. But at the same time it leaves the scalable network in a vulnerable position as no security measures are taken on the cluster. These networks are although cost effective, but the security is next to none. Some companies are using Kerberos mechanism in Hadoop, which is to some extent secures the network using one time random code. It provides node to node or to application authentication before entering the cluster. It also validates MapReduce functionality to identify any untrusted mappers. Layered encryption is also adopted to protect the static data. As it prevents the sensitive information to be directly accessed by administrators or other apps. Key management is also proposed where there is a separate key management server which protects the encrypted keys and also manages separate keys for different files. Apart from service-level approval and web proxy proficiencies from YARN — no proper security service is available to protect data stores or basic Hadoop components [11].

IV. CONCLUSION

The paper presented a survey of well-known security mechanisms used for providing a secure and trusted big data environment. Table 1 presented a summary of Big Data Security Attacks, Challenges, Issues and Methods. A detailed discussion on related security issues and challenges is also carried out. It is clear that the promises big data domain has brought to the field of information technology are immense but can only be fulfilled when proper security mechanisms are in place.

REFERENCES

[1] *Big Data Threat Landscape and Good Practice Guide, Ernesto Damiani (CINI), Claudio Agostino Ardagna*

(CINI), Francesco Zavatarelli (CINI), Evangelos Rekleitis (ENISA), Louis Marinos (ENISA).

[2] *Researchpaper_Unstructured -Big- Data- Processing-Security-Issues-and-Countermeasures, Shivasakthi Nadar, Narendra Gawai, International Journal of Scientific & Engineering Research,* Volume 6, Issue 3, March-2015 201 ISSN 2229-5518,

[3] *Security issues associated with big data in cloud computing. Inukollu, V. N., Arsi, S., & Ravuri, S. R.* (2014). *International Journal of Network Security & Its Applications*, *6*(3), 45.

[4] *Big data security challenges dealing with too many issues, Rashmi N, Uma K M, Jayalakshmi K, Vinodkumar K ,International Journal of Recent Development in Engineering and Technology,* (ISSN 2347-6435(Online) Volume 3, Issue 2, August 2014),

[5]*Top ten security challenges, https://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf*

[6]*Big data security issues,* https://www.alienvault.com/blogs/security-essentials/9-key-big-data-security-issues

[7]*http://www.trendmicro.de/media/wp/addressing-big-data-security-challenges-whitepaper-en.pdf*

[8]*http://www.symantec.com/content/en/us/enterprise/other_resources/bistr_main_report_v19_21291018.en-us.pdf*

[9] *Big-Data Security, Kalyani Shirudkar, Dilip Motwani, Department of Computer Engineering VIT, Mumbai,* India *2012*(7), 5-8.

[10]*https://downloads.cloudsecurityalliance.org/initiatives/bdwg/ Big_Data_Analytics_for_Security_Intelligence.pdf*

[11]*https://securosis.com/assets/library/reports/SecuringBig ata_FINAL.pdf*

[12]*http://www.cisco.com/c/en/us/about/securitycenter/understanding-operational-security.html*