International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Security Challenges Associated with High Dimensional Data

Tata Gayathri*1, N Durga2
1,2Assistant Professor, Department of CSE, Shri Vishnu engineering college for women, Bhimavaram, Andhra Pradesh, India

ABSTRACT

Big data implies performing computation and database operations for massive amounts of data, remotely from the data owner's enterprise. Since a key value proposition of big data is access to data from multiple and diverse domains, security and privacy will play a very important role in big data research and technology. The variety, velocity and volume of big data amplifies security management challenges that are addressed in traditional security management. Big data repositories will likely include information deposited by various sources across the enterprise. This variety of data makes secure access management a challenge. Each data source will likely have its own access restrictions and security policies, making it difficult to balance appropriate security for all data sources with the need to aggregate and extract meaning from the data.

Keywords: Big Data, security, privacy, security Practices

I. INTRODUCTION

Big data implies performing computation and database operations for massive amounts of data, remotely from the data owner's enterprise. Since a key value proposition of big data is access to data from multiple and diverse domains, security and privacy will play a very important role in big data research and technology. The limitations of standard IT security practices are well-known, making the ability of attackers to use software subversion to insert malicious software into applications and operating systems a serious and growing threat whose adverse impact is intensified by big data. So, a big question is what security and privacy technology is adequate for controlled assured sharing for efficient direct access to big data. Making effective use of big data requires access from any domain to data in that domain, or any other domain it is authorized to access. For example, a big data environment may include a dataset with proprietary research information, a dataset requiring regulatory compliance, and a separate dataset with personally identifiable information (PII). In addition, many of the repositories collect data at high volumes and velocity from a number of different data sources, and they all might have their own data transfer workflows. These connections to multiple repositories can increase the attack surface for an adversary. A big data system receiving feeds from 20 different data sources may present an attacker with 20 viable vectors to attempt to gain access to a cluster.

Over the last few years data has become one of the most important assets for companies in almost every field. Not only are they important for companies related to the computer science industry, but also for organisations, such as countries' governments, healthcare, education, ortho engineering sector. Data are essential with respect to carrying out their daily activities, and also helping the businesses' management to achieve their goals and make the best decisions on the basis of the information extracted from them [1]. It is estimated that of all the data in recorded human history, 90 percent has been created in the last few years. In 2003, five exabytes of data were created by humans, and this amount of information is, at present, created within two days [2]. Thistendencytowardsincreasingthevolumeanddetailofth edatathatiscollectedbycompanies will not change in the near future, as the rise of social networks, multimedia, and the Internet of Things (IoT) is producing an overwhelming flow of data [3]. We are living in the era of Big Data. Furthermore, this data is mostly unstructured, signifying that traditional systems are not capable of analysing it. Organisations are willing to extract more beneficial information from this high volume and variety of data [4]. A new analysis paradigm with which to analyse and better understand this data, therefore, emerged in order to obtain not only private, but also public, benefits, and this was Big Data [5].

According to the Big Data Working Group at the Cloud Security Alliance organisation there are, principally, four different aspects of Big Data security: infrastructure security, data privacy, data management, and integrity and reactive security [9]. This division of Big Data security into four principal topics has also been used by the International Organisation for Standardisation in order to create asecurity standard for security in Big Data. Figure1 contains a scheme showing the main topics related to security in Big Data.

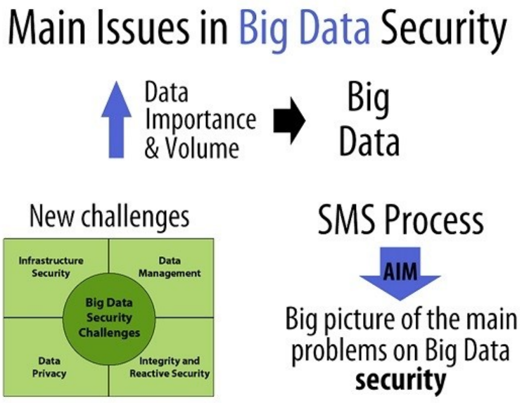
Main Issues in Big Data Security
Data Importance & Volume
Big Data
New challenges
Infrastructure Security
Data Management
Big Data Security Challenges
Data Privacy
Integrity and Reactive Security
SMS Process
AIM
Big picture of the main problems on Big Data security

Figure 1. Main challenges as regards security in Big Data

The purposes of this paper are to highlight the main security challenges that may affect Big Data, along with the solutions that researchers have proposed in order to deal with them. This big picture of the security problem may help other researchers to better understand the security changes produced by the inherent characteristics of the Big Data framework and, consequently, find new research lines so as to carry out more in-depth investigations. This goal has been accomplished by carrying out an empirical investigation by means of the systematic mapping study method with the aim of obtaining a complete background to the security problem as regards Big Data and the proposed solutions.

II. INFRASTRUCTURE

Another big data challenge is the distributed nature of big data environments. Compared with a single high-end database server, distributed environments are more complicated and vulnerable to attack. When big data environments are distributed geographically, physical security controls need to be standardized across all accessible locations. When data scientists across the organization want access to information, perimeter protection becomes important and complicated to ensure access to users while protecting the system from a possible attack. With a large number of servers, there is an increased possibility that the configuration of servers may not be consistent – and that certain systems may remain vulnerable.

The Technology

An additional big data security challenge is that big data programming tools, including Hadoop and NoSQL databases, were not originally designed with security in mind. For example, Hadoop originally didn’t authenticate services or users, and didn’t encrypt data that’s transmitted between nodes in the environment. This creates vulnerabilities for authentication and network security. NoSQL databases lack some of the security features provided by traditional databases, such as role-based access control. The advantage of NoSQL is that it allows for the flexibility to include new data types on the fly, but defining security policies for this new data is not straightforward with these technologies.

III. SECURING BIG DATA

Application Software Security

Use secure versions of open-source software. As described above, big data technologies weren’t originally designed with security in mind. Using open-source technologies like Apache Accumulo or the .20.20x version of Hadoop or above can help address this challenge. In addition, proprietary technologies like Cloudera Sentry or DataStax Enterprise offer enhanced security at the application layer. Specifically, Sentry and Accumulo also support role-based access control to enhance security for NoSQL databases.

Maintenance, Monitoring, and Analysis of Audit Logs.

Implement audit logging technologies to understand and monitor big data clusters. Technologies like Apache Oozie can help implement this feature. Keep in mind that security engineers in the organization need to be tasked with examining and monitoring these files. It's important to ensure that auditing, maintaining, and analyzing logs are done consistently across the enterprise.

Secure Configurations for Hardware and Software. Build servers based on secure images for all systems in your organization's big data architecture. Ensure patching is up to date on these machines and that administrative privileges are limited to a small number of users. Use automation frameworks, like Puppet, to automate system configuration and ensure that all big data servers in the enterprise are uniform and secure.

Account Monitoring and Control. Manage accounts for big data users. Require strong passwords, deactivate inactive accounts, and impose a maximum permitted number of failed log-in attempts to help stop attacks from getting access to a cluster. It's important to note that the enemy isn't always outside of the organization. Monitoring account access can help reduce the probability of a successful compromise from the inside.

IV. SECURITY CHALLENGES

The biggest challenge for big data from a security point of view is the protection of user's privacy. Big data frequently contains huge amounts of personal identifiable information and therefore privacy of users is a huge concern. Because of the big amount of data stored, breaches affecting big data can have more devastating consequences than the data breaches we normally see in the press. This is because a big data security breach will potentially affect a much larger number of people, with consequences not only from a reputational point of view, but with enormous legal repercussions. When producing information for big data, organizations have to ensure that they have the right balance between utility of the data and privacy. Before the data is stored it should be adequately anonymised, removing any unique identifier for a user. This in itself can be a security challenge as removing unique identifiers might not be enough to guarantee that the data will remain anonymous. The anonymized data could be could be cross-referenced with other available data following de-anonymization techniques.When storing the data organizations will face the problem of encryption. Data cannot be sent encrypted by the users if the cloud needs to perform operations over the data. A solution for this is to use "Fully Homomorphic Encryption" (FHE), which allows data stored in the cloud to perform operations over the encrypted data so that new encrypted data will be created. When the data is decrypted the results will be the same as if the operations were carried out over plain text data. Therefore, the cloud will be able to perform operations over encrypted data without knowledge of the underlying plain text data.While using big data a significant challenge is how to establish ownership of information. If the data is stored in the cloud a trust boundary should be establish between the data owners and the data storage owners.Adequate access control mechanisms will be key in protecting the data. Access control has traditionally been provided by operating systems or applications restricting access to the information, which typically exposes all the information if the system or application is hacked. A better approach is to protect the information using encryption that only allows decryption if the entity trying to access the information is authorised by an access control policy. An additional problem is that software commonly used to store big data, such as Hadoop, doesn't always come with user authentication by default. This makes the problem of access control worse, as a default installation would leave the information open to unauthenticated users. Big data solutions often rely on traditional firewalls or implementations at the application layer to restrict access to the information.

Big data is a relatively new concept and therefore there is not a list of best practices yet that are widely recognized by the security community. However there are a number of general security recommendations that can be applied to big data:

- Vet your cloud providers: If you are storing your big data in the cloud, you must ensure that your provider has adequate protection mechanisms in place. Make sure that the provider carries out periodic security audits and agree penalties in case that adequate security standards are not met.
- Create an adequate access control policy: Create policies that allow access to authorized users only.
- Protect the data: Both the raw data and the outcome from analytics should be adequately protected. Encryption should be used accordingly to ensure no sensitive data is leaked.

- Protect communications: Data in transit should be adequately protected to ensure its confidentiality and integrity.
- Use real-time security monitoring: Access to the data should be monitored. Threat intelligence should be used to prevent unauthorised access to the data.

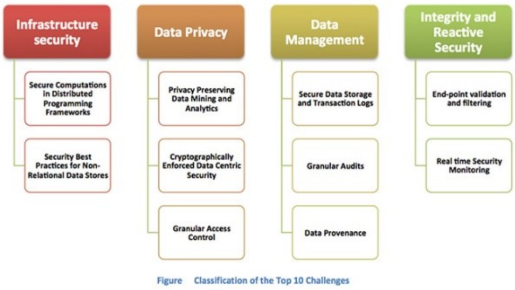
Infrastructure security
Secure Computations in Distributed Programming Frameworks
Security Best Practices for Non-Relational Data Stores
Data Privacy
Privacy Preserving Data Mining and Analytics
Cryptographically Enforced Data Centric Security
Granular Access Control
Data Management
Secure Data Storage and Transaction Logs
Granular Audits
Data Provenance
Integrity and Reactive Security
End-point validation and filtering
Real time Security Monitoring
Figure Classification of the Top 10 Challenges

Figure 2. Classification of security challenges in Big Data security

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 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. Tasks which involve highly parallel computations over large data sets are particularly suited for MapReduce frameworks such as Hadoop. However, the data mappers may contain intentional or unintentional leakages. Untrusted mappers could return wrong results, which will in turn generate incorrect aggregate results. For example, a mapper may emit a very unique value by analyzing a private record, undermining users' privacy. Therefore, there are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper.

Security best practices for non-relational data stores

Relational databases have enjoyed a long run as the database mainstay across a wide variety of businesses, and for good reasons. They're relatively easy to create and use, and they offer reliable performance in both transaction processing and business intelligence applications, with support for transaction and data integrity. However, relational databases haven't necessarily adapted well to changes in the types and quantities of data now being generated, such as the unstructured data that is prevalent in big data applications. In addition, expanding traditional databases to accommodate rapid growth is costly.The proliferation of multiple non-relational databases is transforming the data management landscape. Instead of having to force structures onto their data, organisations can now choose NoSQL database architectures that fit their emerging data needs, as well as combining these new technologies with conventional relational databases to drive new value from their information.

Non-relational databases such as NoSQL are common but they're vulnerable to attacks such as NoSQL injection; the CSA lists a bevy of countermeasures to protect against this. Start by encrypting or hashing passwords, and be sure to ensure end-to-end encryption by encrypting data at rest using algorithms such as advanced encryption standard (AES), RSA, and Secure Hash Algorithm (SHA-256). Transport layer security (TLS) and secure sockets layer (SSL) encryption are useful as well.

Beyond those core measures, plus layers such as data tagging and object-level security, you can also secure non-relational data by using what's called pluggable authentication modules (PAM); this is a flexible method for authenticating users while making sure to log transactions by using a tool such as NIST log. Finally, there's what's called fuzzing methods, which expose cross-site scripting and injecting vulnerabilities between NoSQL and the HTTP protocol by using automated data input at the protocol, data node, and application levels of the distribution.

Secure Data Storage and Transaction Logs Storage management is a key part of the Big Data security equation. The CSA recommends using signed message digests to provide a digital identifier for each digital file or document, and to use a technique called secure untrusted data repository (SUNDR) to detect unauthorized file modifications by malicious server agents.

The handbook lists a number of other techniques as well, including lazy revocation and key rotation, broadcast and policy-based encryption schemes, and digital rights management (DRM). However, there's no substitute for simply building your own secure cloud storage on top of existing infrastructure. Endpoint Filtering and Validation Endpoint security is paramount and your organization can start by using trusted certificates, doing resource testing, and connecting only trusted devices to your network by using a mobile device management (MDM) solution (on top of antivirus and malware protection software). From there, you can use statistical similarity detection techniques and outlier detection techniques to filter malicious inputs, while guarding against Sybil attacks (i.e., one entity masquerading as multiple identities) and ID-spoofing attacks.

Real-Time Compliance and Security Monitoring Compliance is always a headache for enterprises, and even more so when you're dealing with a constant deluge of data. It's best to tackle it head-on with real-time analytics and security at every level of the stack. The CSA recommends that organizations apply Big Data analytics by using tools such as Kerberos, secure shell (SSH), and internet protocol security (IPsec) to get a handle on real-time data.Once you're doing that, you can mine logging events, deploy front-end security systems such as routers and application-level firewalls, and begin implementing security controls throughout the stack at the cloud, cluster, and application levels. The CSA also cautions enterprises to be wary of evasion attacks trying to circumvent your Big Data infrastructure, and what's called "data-poisoning" attacks (i.e., falsified data that tricks your monitoring system).

Preserve Data Privacy
Maintaining data privacy in ever-growing sets is really hard. The CSA said the key is to be "scalable and composable" by implementing techniques such as differential privacy—maximizing query accuracy while minimizing record identification and homomorphic encryption to store and process encrypted information in the cloud. Beyond that, don't skimp on the staples: The CSA recommends incorporating employee awareness training that focuses on current privacy regulations, and being sure to maintain software infrastructure by using authorization mechanisms. Finally, the best practices encourage implementing what's called "privacy-preserving data composition," which controls data leakage from multiple databases by reviewing and monitoring the infrastructure that's linking the databases together.

BigData Cryptography

Mathematical cryptography hasn't gone out of style; in fact, it's gotten far more advanced. By constructing a system to search and filter encrypted data, such as the searchable symmetric encryption (SSE) protocol, enterprises can actually run Boolean queries on encrypted data. After that's installed, the CSA recommends a variety of cryptographic techniques.

Relational encryption allows you to compare encrypted data without sharing encryption keys by matching identifiers and attribute values. Identity-based encryption (IBE) makes key management easier in public key systems by allowing plaintext to be encrypted for a given identity. Attribute-based encryption (ABE) can integrate access controls into an encryption scheme. Finally, there's converged encryption, which uses encryption keys to help cloud providers identify duplicate data.

Granular Access Control

Access control is about two core things according to the CSA: restricting user access and granting user access. The trick is to build and implement a policy that chooses the right one in any given scenario. For setting up granular access controls, the CSA has a bunch of quick-hit tips:

- Normalize mutable elements and denormalize immutable elements,
- Track secrecy requirements and ensure proper implementation,
- Maintain access labels,
- Track admin data,
- Use single sign-on (SSO), and
- Use a labeling scheme to maintain proper data federation.

Granular Audit

Granular auditing is a must in Big Data security, particularly after an attack on your system. The CSA recommends that organizations create a cohesive audit view following any attack, and be sure to provide a full audit trail while ensuring there's easy access to that data in order to cut down inciden t response time.

Audit information integrity and confidentiality are also essential. Audit information should be stored separately and protected with granular user access controls and regular monitoring. Make sure to keep your Big Data and audit data separate, and enable all required logging when you're setting up auditing (in order to collect and process the most detailed information possible). An open-source audit layer or query orchestrator tool such as Elastic Search can make of all this easier to do.

Data Provenance

Data provenance can mean a number of different things depending on who you ask. But what the CSA is referring to is provenance metadata generated by Big Data applications. This is a whole other category of data that needs significant protection. The CSA recommends first developing an infrastructure authentication protocol that controls access, while setting up periodic status updates and continually verifying data integrity by using mechanisms such as checksums.

V. FUTURE SCOPE AND DEVELOPMENT

As far as the future of big data is concerned it is for certain that data volumes will continue to grow and the prime reason for that would be the drastic increment in the number of hand held devices and internet connected devices, which is expected to grow in an exponential order. Machine learning will have a far bigger role to play for data preparation and predictive analysis in businesses in the coming days. Privacy and security challenges related to big data will grow and by 2018, 50% of business ethics violations will be related to data.

VI.CONCLUSION

We represented “Big data Security”. Big data have various challenges related to security like-computation in distributed programming, security of data storage and transaction log, input filtering from client, scalable data mining and analytics, access control and secure communication. For tackling with such security challenges we used different security methods like Type Based keyword search for security of big data, use of hybrid cloud to provide privacy in big data . As far as security is concerned the existing technologies are promising to evolve as newer vulnerabilities to big data arise and the need for securing them increases.

VII. REFERENCES

[1]. Mayer-Schonberger, V.; Cukier, K. BigData:RevolutionthatWillTransformHowWeLive,Work,andThink; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.

[2]. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20-24 May 2013; pp. 42-47.

[3]. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Ullah Khan, S. The rise of "big data" on cloud computing: Review and open research issues. Inf. Syst. 2015, 47, 98-115. [CrossRef]

[4]. Sharma, S. Rise of Big Data and related issues. In Proceedings of the 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 17-20 December 2015; pp. 1-6.

[5]. Eynon, R. The rise of Big Data: What does it mean for education, technology, and media research? Learn. Media Technol. 2013, 38, 237-240. [CrossRef]

[6]. Wang, H.; Jiang, X.; Kambourakis, G. Special issue on Security, Privacy and Trust in network-based Big Data. Inf. Sci. Int. J. 2015, 318, 48-50. [CrossRef]

[7]. Thuraisingham, B. Big data security and privacy. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, San Antonio, TX, USA, 2-4 March 2015; pp. 279-280.

[8]. Rijmenam, V. ThinkBigger: DevelopingaSuccessful Big DataStrategyfor YourBusiness; Amacom: New York, NY, USA, 2014.

[9]. Big Data Working Group; Cloud Security Alliance (CSA). Expanded Top Ten Big Data Security and Privacy. April 2013. Available onlinehttps://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_ Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf (accessed on 9 December 2015).

[10]. Meng, X.; Ci, X. Big data management: Concepts, techniques and challenges. Comput. Res. Dev. 2013, 50, 146-169.

[11]. Xu, L.; Jiang, C.; Chen, Y.; Ren, Y.; Liu, K.J.R. Privacy or Utility in Data Collection? A Contract Theoretic Approach. IEEE J. Sel. Top. Signal Proc. 2015, 9, 1256-1269.

[12]. Cheng, H.; Rong, C.; Hwang, K.; Wang, W.; Li, Y. Secure big data storage and sharing scheme for cloud tenants. China Commun. 2015, 12, 106-115. [CrossRef]

[13]. Weber, A.S. Suggested legal framework for student data privacy in the age of big data and smart devices. In Smart Digital Futures; IOS Press: Washington, DC, USA, 2014; Volume 262.

[14]. Thilakanathan, D.; Calvo, R.; Chen, S.; Nepal, S. Secure and controlled sharing of data in distributed computing. In Proceedings of the 16th IEEE International Conference on Computational Science and Engineering (CSE 2013), Sydney, Australia, 3-5 December 2013; pp. 825-832.

[15]. Chen,J.;Liang,Q.;Wang,J.Securetransmissionforbigdatabasedonnestedsamplingandcoprimesampling with spectrum efficiency. Secur. Commun. Netw. 2015, 8, 2447-2456. [CrossRef]

[16]. Liu, C.; Yang, C.; Zhang, X.; Chen, J. External integrity verification for outsourced big data in cloud and IoT. Future Gener. Comput. Syst. 2015, 49, 58-67. [CrossRef]

[17]. Wang, Y.; Wei, J.; Srivatsa, M.; Duan, Y.; Du, W. IntegrityMR: Integrity assurance framework for big data analytics and management applications. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6-9 October 2013; pp. 33-40.

[18]. Liao, C.; Squicciarini, A. Towards provenance-based anomaly detection in MapReduce. In Proceedings of the 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Shenzhen, China, 4-7 May 2015; pp. 647-656.

[19]. Tan, Z.; Nagar, U.T.; He , X.; Nanda, P.; Liu, R.P.; Wang, S.; Hu, J. Enhancing big data security with collaborative intrusion detection. IEEE Cloud Comput. 2014, 1, 27-33. [CrossRef]

[20]. Chang, V. Towards a Big Data system disaster recovery in a Private Cloud. Ad Hoc Netw. 2015, 35, 65-82. [CrossRef]

[21]. J.Z. Huang, M.K. Ng, H. Rong, and Z. Li, "Automated Variable Weighting in k-Means Type Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 1-12, May 2005.

[22]. L. Jing, M.K. Ng, J. Xu, and J.Z. Huang, "Subspace Clustering of Text Documents with Feature Weighting k-Means Algorithm," Proc. Ninth Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp. 802-812, 2005.

[23]. L. Parsons, E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A Review," SIGKDD Explorations, vol. 6, no. 1, pp. 90-105, 2004.

[24]. C.H. Cheng, A.W. Fu, and Y. Zhang, "Entropy-Based Subspace Clustering for Mining Numerical Data," Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge and Data Mining, pp. 84-93, 1999.

[25]. S. Goil, H. Nagesh, and A. Choudhary, "Mafia: Efficient and Scalable Subspace Clustering for Very Large Data Sets," Technical Report CPDC-TR-9906-010, Northwest Univ., 1999.

[26]. R.T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Int'l Conf. Very Large Data Bases, pp. 144-155, Sept. 1994.

[27]. G. De Soete, "Optimal Variable Weighting for Ultrametric and Additive Tree Clustering," Quality and Quantity, vol. 20, pp. 169180, 1986.

[28]. Big Data: Issues and Challenges Moving Forward. 2013 46th Hawaii International Conference on System Sciences Stephen