Enforcing Data Security in A Cloud based Big Data Environment- A Critical Study

Dr. Lakshmi Priya Vinjamuri
Associate Professor, Law College Dehradun, Uttaranchal University

Ms. Aghanaashaa. A
8th Semester B.Tech (IT), Chandigarh University

ABSTRACT:
Data security in cloud computing is an emerging and intensive area that is highly vulnerable to various glitches and violations. The big data environment which is predominantly cloud based has confidential and data that mandates secrecy of high order owing to the sensitive nature of the information that needs data security measures that are not only robust but are also effective, efficient and economical.
Data security and privacy is an emerging area of information security and technology that requires a stringent protocol of security administration and a parallel legal framework that addresses the challenges of data violation and privacy infringement.
The paper is an attempt to study the enforcement of data privacy and security protocols in a cloud computing environment. The paper is an insight into the data security technologies, the various kinds of cryptography and addresses the need for a comprehensive security framework with respect to privacy and security while dealing with big data in a cloud environment.

Keywords: Data security, privacy, security protocols, cryptography, data security technologies

INTRODUCTION:

Big Data and Cloud Computing:

'Bigdata'dealswithmassivestructured, semi structuredorunstructureddatato storeandprocess fordata-analysispurpose.

There are five aspects of Big Data which are described through 5Vsnamely volume, variety, velocity, value and veracity.

Volume of the data signifies the amount of the data, variety the different types of data, value indicated the value of the data based on the information contained within and veracity signifies the data confidentiality and availability.

Cloud computing off ers services to the users on a pay-as-you go model and the cloud providers offer three primary services that include infrastructure as a service denoted by IAAS[1] where the service provider offers entire infrastructure along with the maintenance related tasks, the platform as a service depicted as PAAS[2] in which the resources like object storage, runtime,

[1]https://www.gartner.com/en/information-technology/glossary/infrastructure-as-a-service-iaas#:~:text=Infrastructure%20as%20a%20service%20(IaaS)%20is%20a%20standardized%2C%20highly,time%20and%20metered%20by%20use.

[2]https://www.techtarget.com/searchcloudcomputing/definition/Platform-as-a-Service-PaaS#:~:text=Platform%20as%20a%20service%20(PaaS,software%20on%20its%20own%20infrastructure.

queuing, databases are provided and the responsibility of configuration and implementation related tasks depend on the consumer and the software as a service commonly known as SAAS which is highly facilitated as it provides all the necessary setting and an infrastructure is predesigned for the platform

Cloudcomputingandbigdata are anidealcombinationas a comprehensive solution on which both scalable and accommodation for the big data and business analytics are provided together.

The salient features of big data and cloud computing are agility, elasticity, data processing, cost-cutting with big data on the cloud. Agility in cloud computing facilitates the infrastructure with all the required resources almost instantly and a good cloud provider will ensure that work is always on the go without any hitches in contrast to the traditional method of storing and manging the data which is expensive, time-consuming with a long server set-up haul.

The second aspect of elasticity can be understood on a cloud platform to be a phenomenon that the cloud platform can dynamically expand to provide storage for ever increasing data and once the company or organization gets the necessary insight from the data storage space can be increased or reduced to accommodate the data as per the requirement.

Data processing involves a large volume of data that leads to the issue of how to process it efficiently and social media alone generates massive amount of unstructured data in various forms and with big data platforms cloud computing makes the process not only easier but also accessible to small, medium and large enterprises.

Cutting of costs with big data in the cloud is the fourth and crucial aspect in cloud computing which is a terrific solution for enterprises that wish to have state-of-the- art technology running their operations under limited budget constraints. Maintaining a big data centre to perform big data analytics can quickly drain an IT budget and nowadays companies have the option to avoid investing in setting up the IT department and maintaining hardware infrastructure.Cloud computing facilitates the responsibility shift to the cloud providers and the company only has to pay for the storage space and power consumption.

Reduced complexity is possible with cloud computing. Any implementation of big data solution requires several components and integrations and cloud computing provides the option to automate these components thereby reducing the complexity and enhancing the productivity of big data analytics team.

The cloud based big architecture [4]is depicted in image 1.1 below for an understanding of what cloud computing is capable of doing especially with big data, its analytics and security.

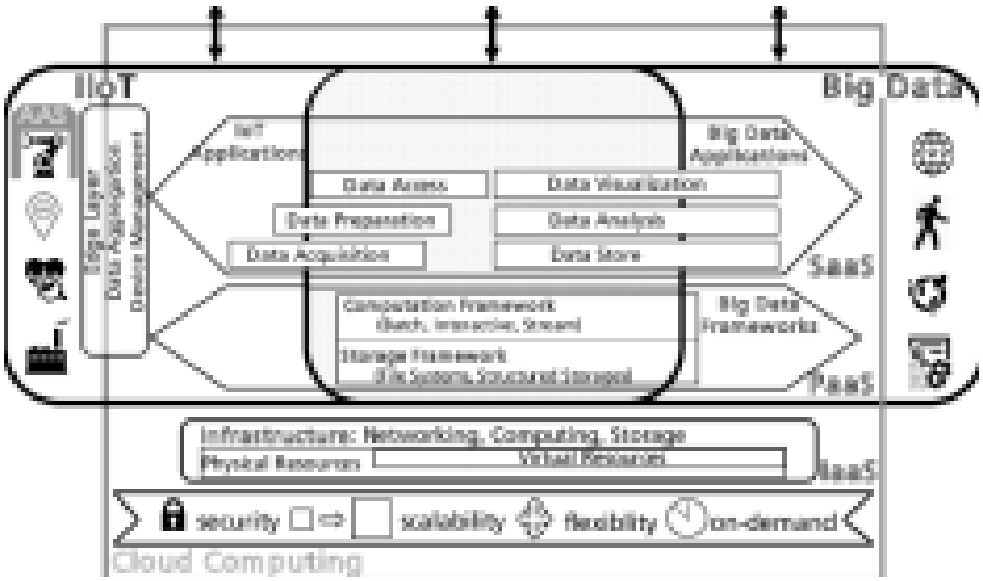
DATASECURITY:

[3]https://www.computer.org/publications/tech-news/trends/big-data-and-cloud-computing

[4]https://www.vmware.com/topics/glossary/content/cloud-architecture.html

Prioritizing big data security low and putting it off till later stages of big data adoption in projects isn't always a smart move. People don't say "security first" for no reason. At the same time, we admit that ensuring big data security comes with its concerns and challenges which is why it is more than helpful to get acquainted with them.

Some of the major challenged of securing data include authentication level issues, data level issues and network level issues.



Cloud based Big Data architecture

Authentication level issues:

There are many clusters of data present on the cloud and on big data and every node has a different priority of access or rights to function, nodes with administrative rights have the major access to any data. However, if malicious software gets into a node of administrative priority, then there is a risk of stealing theft and manipulating the critical user data. Forfasterexecutionwith parallel processing many nodes join into huge clusters.

In case of no authentication any malicious node can disturb the cluster. Logging plays an important role in big data. If logging is not provided then no activity is recorded which modifies or deletes the data. If new node joins the cluster, then that will not be recognized because of logging absence.

Data level issues:

Data is very important part and also plays vital role. Data is nothing but raw facts and fig ures that areusedbytheusersonanyplatformwhichmayincludesocialnetworkingsites, reservationsites,governmentsitesor accessingdatafromvariousportals.Datalevelissuesdealwithdataintegrityand availability such as data protection, privacy, safety and distributed data. To improve efficiency many big data environments like Hardtop store the data is as it is without data encryption. If the hacker accesses the machines, then it is impossible to stop him. In distributed data store, information is

stored in many nodes with replicas but if any of the replica data is attacked it becomes next to impossible to recover the data.

Network Level Issues:

There are many nodes present in clusters and computation or processing of data is done in these nodes. This processing of data can be done anywhere among the nodes in clusters. So, it is difficult to find on which node data is processing because of this difficulty on which node security would be provided is often complicated. Two or more nodes can communicate with each other or share their data/resources through network.

General Level Issues

Many technologies are used for processing the data and some traditional security tools for security purposes. Traditional tools are developed over years ago. So, these tools may not be performed well with new distributed form of big data. As big data uses many technologies for data storing, data processing and data retrieval, there may be some complexities that occur because of the varied technology.

Secure Computations

Bid data technologies use distributed programming frameworks to process large amounts of data. These distributed frameworks like map reduce don't have good security protections. In map reduce, the data is split, then processed by a mapper and allocated storage. If someone can change the mapper settings as it doesn't have any additional security layer, it can manipulate the data being processed.

Also, it is very difficult to detect these not trusted mappers. It is very important to secure the computations being handled in these distributed programming frameworks so as to ensure that the integrity of the data is maintained.

Protecting Data and Transaction logs:

Due to the size of data and transaction logs,these are stored in multi-tiered storage environments with auto tiering functionality. Auto tiering does not keep track of the data location. Auto tiering systems can expose new vulnerabilities because of unknown physical data locations, non-trusted storage devices which can result in organizations losing control over their data. Data transmission between tiers can also provide information regarding user activities and data properties which can be used by attackers. Data transaction logs need to maintain the confidentiality, integrity and availability of data.

Validation of Inputs from Endpoints:

Big data collects data from a variety of input devices including endpoints. It may be collecting logs from a large number of devices and applications. The data which big data is receiving might contain rouge data being sent by a non-trusted endpoint. This can affect the organisation's analytical outputs. A challenge here is to validate all the inputs the big data is receiving to ensure that it came from a trusted source.

Secure Non-Relational Data Stores:

Non-Relational data stores like NoSQL are rapidly being used in big data technologies. These data stores are not mature enough, as of today. They have many security issues like no encryption support for data files, weak authentication between client and server, data at rest is unencrypted which can cause privacy threats.

Privacy-preserving data analysis:

Privacy is an important issue in applying big data technologies for analytics. As more and more data are being collected, the data aggregation along with data analytics could result in user privacy violation. If the data analytics is outsourced, an untrusted third-party employee can infer personal information of users. The organizations want to use big data analytics tools to ensure protecting user privacy while doing so.

Access control:

Big data handles a variety of data including sensitive data such as personally identifiable information of users. There are many legal and compliance requirements to protect the data. Granular access control policies should be implemented so that only authorized users to have access to sensitive user data and analytics done on those data sets. This is needed to ensure confidentiality of data.

Real-time security monitoring:

Real time security monitoring is needed for big data infrastructure and the analytics it is handling. It has always been a difficult task because of the number of alerts generated by devices. These alerts have a large number of false positives as well. Due to this reason, companies often struggle to monitor real-time data.

Criteria for Data Access (CIA Traid[5])

a) Availability can be defined as the attestation or guarantee that data will be available to the user in a perpetual manner irrespective of the location or point of access of the user. It is ensured by: fault tolerance, network security and authentication.
b) Integrity is the assurance that the data sent is same as the message received and it is not altered in between. Integrity is infringed if the transmitted message is not same as received one. It is ensured by: firewalls and intrusion detection.
c) Confidentiality is the avoidance of unauthorized exposure of user data. It is ensured by: security protocols, authentication services and data encryption services.

Data Security Technologies

Encryption: Your encryption tools need to secure data in-transit and at rest, and they need to do it across massive data volumes. Encryptions also needs to operate on many types of data, both user and machine generated. Encryption tools also need to work with different analytics toolsets and their output data, and on common big data storage formats including relational database

[5] https://www.f5.com/labs/articles/education/what-is-the ciatriad#:~:text=These%20three%20letters%20stand%20for,objectives%20for%20every%20security%20program.

management systems (RDBMS), non-relational databases like nasal and specialized files systems such as Hadoop Distributed File System (HDFS[6])

Centralized Key Management: Centralized Key Management has been a security best practice for many years. It applies just as strongly in big data environments, especially those with wise geographical distribution. Best practices include policy driven automation, logging, on demand key delivery and abstracting key management from key usage.

User Access Control: User access control may be the most basic network security tool, but many companies practice minimal control because the management overhead can be very high. This is dangerous enough at the network level, and can be disastrous for the big data platform. Strong user access control requires a policy-based approach that automated access based on user and role-based settings, Policy driven automation manages complex user control levels, such as multiple administrator settings that protect the big data platform against inside attacks.

Intrusion Detection and Prevention: Intrusion detection and prevention systems are security workhouses. This does not make them less valuable to the big data platform. Big data's value and distributed architecture lends itself to intrusion attempts. IPSenables security admins to protect the bid data platform from intrusion, and should an intrusion succeed, IDS quarantines the intrusion before it does significant damage.

Physical Security: Do not ignore physical security. Build it in when you deploy your big data platform in your own data centre, or carefully do due diligence around your cloud provider's data centre security. Physical security systems can deny data centre access to strangers or to staff members who have no business being in sensitive areas. Video surveillance and security logs will do the same.

Types of Cryptography:

In general, there are two types of cryptography:

a) Symmetric Key Cryptography
b) Asymmetric Key Cryptography

Symmetric Key Cryptography[7]:

It is an encryption system where the sender and receiver of a message that use a single common key to encrypt and decrypt messages. Symmetric Key Systems are faster and simpler but the problem is that sender and receiver have to somehow exchange keys in a secure manner. The most popular symmetric key cryptography system is Data Encryption System (DES). Symmetric encryption is a type of encryption whereby only one cryptographic key(secret key) is used to both encrypt and decrypt electronic information.

The entities communicating through symmetric encryption must exchange the key so that it can be used in the decryption process.

[6] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[7] https://www.hypr.com/symmetric-key-cryptography/#:~:text=Symmetric%20Key%20Cryptography%20also%20known,another%20is%20used%20to%20decrypt.

When a server and client are in need of a secure encrypted message, they send a query over the network to the other party which eventually sends back a copy of the certificate. The other party's cryptographic public key can be extracted from the certificate. The commons symmetric encryption algorithm include:RC4,AES,DES,3DES and, QUAD.

Asymmetric Key Cryptography[8]:

Under this system a pair of keys is used to encrypt and decrypt information. A public key is used for encryption and a private key for decryption. Public key and Private Key are different, even if the public key is known by everyone the intended receiver can only decode it because he alone knows the private key. Asymmetric encryption also referred to as public key cryptography, is a type of encryption whereby two cryptography keys are used to encrypt a plaintext. Secret keys (one public and another private) are exchanged over the internet or a large network. It ensured that malicious persons do not misuse the keys.
The first public key is made freely available to anyone who might want to send you a message whereas the second private key is kept a secret so that you can only know. Here, a message that is encrypted using private key can only be decrypted using a public key while a message that is encrypted using a public key can only be decrypted using a private key. The most common asymmetric key encryption algorithms include PKCS, Elliptic curve techniques, RSA and DES.

Algorithms of Asymmetric Keys:

Diffie-Hellman[9]: The first prime-number, security key algorithm was named Diffie Hellman algorithm and patented in 1977. The Diffie-Hellman algorithm [10]is non authenticated, but does require sharing of a "secret" key between the two communicating parties. The two parties agree on an arbitrary starting number that they share, then each selects a number to be kept private.

In the critical exchange, each party multiplies their secret number by the public number, and then they exchange the result. When each multiplies the exchanged numbers with their private numbers, the result should be identical providing provenance between the parties. It is difficult, computationally party listener to derive the private numbers.
However, in the absence of authentication, Diffie Hellman is vulnerable to man-in the middle attacks, where the third party can interpret communications appearing as a valid participant in the communication while changing or stealing information.

Rivest Shamir Adleman(RSA):[11]RSA, which is patented in 1983 and still the most widely used system for digital security, was released the same year as Diffie Hellman, and was named after its inventors, Ron Rivest, Adi Shamir, and Leonard Adelman. RSA gets much of its added security by combining two algorithms: one is applied to asymmetric cryptography, or PKI (Public Key

[8]https://www.techtarget.com/searchsecurity/definition/asymmetric-cryptography

[9]https://www.hypr.com/diffie-hellman-algorithim/#:~:text=The%20Diffie%E2%80%93Hellman%20(DH)%20Algorithm%20is%20a%20key%2D,or%20data%20using%20symmetric%20cryptography.

[10] Ibid.

[11]https://www.techtarget.com/searchsecurity/definition/RSA

Infrastructure), and the other algorithm provides for secure digital signatures. While the essential mathematics of both components is similar, and the output keys are of the same format.

The RSA algorithm has three main processes: key pair generation, encryption and decryption. Key pairs include the generation of the public key and the private key. because of this part of the process, RSA has often been described as the public- key digital security system. Once the public key is generated, it is transmitted over an unsecured channel, but the private key remains secret and is not shared with anyone. The data is encrypted with the public key, but can only be decrypted with the private key. The keys are generated by multiplying large prime numbers. Since, as we noted, it was fast and easy to multiply ever larger numbers, prime number encryption became a standard through several decades. To add a layer of security a method of obtaining digital signatures was an additional improvement in RSA. In this scenario-to simplify the process-the sender produces a hash value of the message, which uses the same exponentiation as the encryption number. The receiver does the same hash value at the receiving end to arrive at the same number, confirming the secured signature.

Other protocols rely on RSA has had a lot of staying power in the security world as other certification and security schemes have piggybacked onto it. However, RSA digital signature has a vulnerability, which will result in brute force attacks being able to decode the private key, and exposed to specific attack types as side channel analysis, timing attacks, and others.

In addition, there is computational overhead involved in RSA, and particularly in mobile and table environment, as a result, the performance issue is a great deal. Key length is also a concern, as RSA keys now must be 2048-bit long, because given advances in cryptography and computing resources, 1024-bit keys were deemed insufficiently secure against several attacks. Government and many other organizations are now requiring a minimum key length of 2048-bits.

a) Digital Signature Algorithm(DSA):[12]

In 1991, the National Security Agency (NSA) developed the Digital Signature Algorithm (DSA) as an alternative to the RSA algorithm. The National Institute of Standards and Technology (NIST) gave the algorithm its sanction as US government approved and certified encryption scheme that offered the same degree of security as RSA, but employs different mathematical algorithms for signing and encryption.

Like RSA, DSA is an asymmetric encryption scheme, or PKI, which generates a pair of keys, one public and one private. The signature is created privately, though it can be identified publicly, the benefit of this is that only one authority can create the signature, but any other party can validate the signature using the public key. DSA, as a result, is faster in signing, but slower in verifying, hence, DSA is a sensible choice if there are more performance issues on the client side. DSA and RSA can be run together under some server systems like Apache, providing additional protection.

However, being so similar, DSA and RSA are subject to similar attacks, and RSA has moved to longer keys, which DSA has not yet done. While creating longer DSA key is theoretically possible, it is not being done, so despite having very comparable in other ways to RSA, RSA remains the preferred encryption scheme.

[12] https://www.simplilearn.com/tutorials/cryptography-tutorial/digital-signature-algorithm

CONCLUSION

The research critically establishes the use of algorithms of various kinds in the encryption algorithm family to protect the confidentiality, integrity and availability of big data in a cloud-based environment. The different options for data protection have been analysed based on the scheme and requirement as required by the client or organisations which handle big data in a cloud. The paper offers a future scope to identify the ideal and best algorithm that can be used to meet the requirements of data security, confidentiality and privacy on the cloud.

REFERENCES:

[1]. https://www.gartner.com/en/information-technology/glossary/infrastructure-as-a-service-iaas#:~:text=Infrastructure%20as%20a%20service%20(IaaS)%20is%20a%20standardized%2C%20highly,time%20and%20metered%20by%20use.
[2]. https://www.techtarget.com/searchcloudcomputing/definition/Platform-as-a-Service-PaaS#:~:text=Platform%20as%20a%20service%20(PaaS,software%20on%20its%20own%20infrastructure
[3]. https://www.computer.org/publications/tech-news/trends/big-data-and-cloud-computing
[4]. https://www.vmware.com/topics/glossary/content/cloud-architecture.html
[5]. https://www.f5.com/labs/articles/education/what-is-the-cia triad#:~:text=These%20three%20letters%20stand%20for,objectives%20for%20every%20security%20program.
[6]. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
[7]. https://www.hypr.com/symmetric-key-cryptography/#:~:text=Symmetric%20Key%20Cryptography%20also%20known,another%20is%20used%20to%20decrypt.
[8]. https://www.techtarget.com/searchsecurity/definition/asymmetric-cryptography
[9]. https://www.hypr.com/diffie-hellman-algorithim/#:~:text=The%20Diffie%E2%80%93Hellman%20(DH)%20Algorithm%20is%20a%20key%2D,or%20data%20using%20symmetric%20cryptography.
[10]. https://www.techtarget.com/searchsecurity/definition/RSA
[11]. https://www.simplilearn.com/tutorials/cryptography-tutorial/digital-signature-algorithm
[12]. www.researchgate.net/publication/283748968_ECC_based_image_encryption_scheme_with_aid_of_optimization_technique_using_differential_evolution_algorithm
[13]. https://dmg.tuwien.ac.at/drmota/koppens teinerdiplomarbeit.pdf
[14]. https://andrea.corbellini.name/2015/05/17/elliptic-curve-cryptography-a-gentle introduction/
[15]. https://hackernoon.com/what-is-the_math-behind-elliptic-curve cryptography-f61b25253da3
[16]. www.hindawi.com/journals/scn/2019/46 56281/#introduction
[17]. https://www.geeksforgeeks.org/cryptography-and-its-types/
[18]. https://www.khanacademy.org/computing/computer-science/cryptography
[19]. https://www.edureka.co/blog/what-is cryptography/:

[20]. https://crypto.stackexchange.com/questions/10097/elliptic-curve-cryptography encryption-results
[21]. https://searchsecurity.techtarget.com/definition/asymmetric-cryptography
[22]. https://www.researchgate.net/figure/Cloud-based-Industrial-IoT-and-Big-Data Architecture-15-38_fig5_326539673
[23]. scihub.tw/https://link.springer.com/article/ 10.1007/s10586-017-1542-8
[24]. https://www.google.com/search?q=elliptic+curve+digital+signature+algorithm&oq=elliptical+cure&aqchrome.3.69i57j0 l7.56681j7&sourceid=chrome&i.e.,=TF-8
[25]. https://cse.iitkgp.ac.in/~abhij/download/doc/ECC.pdf