A STUDY ON SECURITY OF BIG DATA STORED ON CLOUD PLATFORMS

Adhitya Rajagopalan[1]

[1]*Student, Department of Computer Science, BITS Pilani, Dubai Campus*

Abstract

Big Data is a field of work that is growing on a daily basis. There is a need to understand what it means, and its interpretation from a technical perspective. An insight into this could be provided by studying the Philosophy of Big Data [1]; the definitions and meanings regarding data that is 'big' in volume, velocity and variability. Cloud Computing is an upcoming service/platform that allows users to store their data online and reduces the need to invest in physical infrastructure that would be necessary to store the data locally. As the need for Big Data increases, so does the necessity to keep it secure. Data security, especially on the cloud, is a major challenge posed to specialists and the development of a cloud service that understands the threats posed to it and take counter-active measures to ensure the security of data stored on it is of utmost importance. This paper analyzes what Big Data and Cloud Computing are about, how large volumes of data are stored on the cloud platform and the security measures that are being followed, the encryption schemes like the Advanced Encryption Standard (AEN), Homomorphic Encryption, and Attribute Based Encryption.

Keywords: *Big Data, Cloud Computing, Encryption Schema, Data Security*

--***--

1. INTRODUCTION

Data is an integral part of any business. Data plays an important role in decision making. Often, decisions made without considering the 'Data Factor' tend to disrupt the normal work flow of an organization. Data is referred to as facts and statistics collected together for an analysis of any given problem/ experiments. One of the newest and by-far the strongest presence in everyday life is **Big Data**. These days, data is everywhere- ranging from scientific data to corporate data to commercial data to personal data. **Data Science** is becoming a very important field of study and involves big data, the theories surrounding it, the statistical measures used to achieve these theories/ proposals and information technology. The term **information** generally refers to the effort put in by human beings to measure the extensive properties of human knowledge.

Now-a-days computing is being transformed into a cluster of services that can be commoditized and delivered to the end user in a way that is similar to utilities such as electricity, water etc. These services can be accessed as per the requirements of the end user, regardless of where these services are hosted. One of the ways in which this is implemented is through **cloud computing.** Cloud Computing is based on dynamic provisioning [2]. As a result of this, almost everyone is able to create a sustainable back-end to host an application and its servers. This way, consumers do not have to invest on building and maintaining IT infrastructure and instead focus solely on creating their commodity. The term **cloud** denotes the infrastructure from which businesses and users can access applications as services from anywhere in the world and on demand. The two main features of cloud computing that make it unique are: availability and flexibility. The major issue with cloud computing is that the owner of the data does not have control over the data at all times. This raises a major security concern as far as cloud computing is concerned. A major challenge that is being faced by cloud specialists is the development of a security protocol for storage, sharing and retrieval of data. With the volume of data that is getting added on to the cloud every minute, it is the duty of the cloud provider to ensure the safety and availability of customer data at all times.

Cloud computing services, with the ability to ingest, store and analyze big data are becoming popular and new models are being developed daily to overcome the challenges associated with Big Data and cloud computing. Data Security and privacy is very important as far as storage of Big Data in clouds are concerned. To overcome these shortcomings, a number of **encryption** techniques (AES, Homomorphic Encryption, Attribute Based Encryption etc.) have been proposed and various methods have been developed to secure data storage, transfer and access on the cloud. However, most of these techniques are not efficient. A better and more secure model would be a **Neural Data Security Model.** This would ensure better data confidentiality and security in cloud environments. The security component of this model would be achieved using the RSA encryption algorithm, and dynamic hashing ensures the confidentiality component.

2. EVOLUTION OF BIG DATA

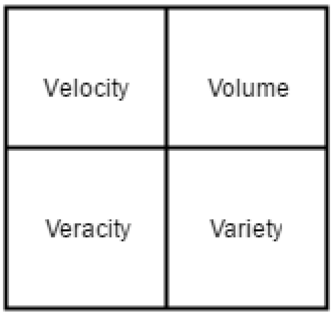
Big data is high-volume, high velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. [3] There are four parameters through which big data is usually analyzed:

i) **Velocity**: This refers to the speed at which the data is generated, or the frequency with which it is being delivered. This is especially important in sensitive processes such as catching fraud. To maximize the value of data in such cases, big data must be used as it streams into the application in which it is being used.

ii) **Variety**: Data from a vast variety of sources are being accumulated daily. There are increasing number of file types that are being uploaded daily, and each of these file types contain different types of data.

iii) **Volume**: It refers to size of data set that we are dealing with. Big Data, as the name suggests means that a huge data set is under consideration.

iv) **Veracity**: This fourth parameter is often overlooked, but is just as important as the other three factors. Veracity refers to the data in doubt, i.e., it refers to the inconsistency, incompleteness, ambiguity, latency and deception of data under scrutiny.



Velocity	Volume
Veracity	Variety

Fig -1: Dimensions of Big Data

3. BIG DATA CHALLENGES

Since there is not enough storage capacity to handle massive amounts of data, big data scientists are faced with a variety of challenges

A. **Security and Privacy**: The data that we collect for our use must be dependable and must be gathered from reliable sources. Data sets consisting of very sensitive data give rise to countless possibilities for misuse and unauthorized access.

B. **Improper Usage of Big Data**: A lot of companies generate large amounts of data, but very few companies are actually making use of this data that is being generated. An in-depth analysis of this data could increase the productivity of the organization.

C. **Natural Language Processing and Machine Learning Development**: Machine learning is a field of work that is developing rapidly. For this area of work to reach its fullest potential, fast and efficient big data analysis and processing techniques must be developed.

D. **Data Search**: Finding high-quality data from a large volume of data are a challenge posed to Big Data Scientists. Pinpointing one stream of data that is important from an otherwise useless data ocean is extremely difficult. Search algorithms that work with normal small-stream data would not work/ be inefficient with big-data.

E. **Quality of Data**: Finding a piece of data that is relevant to a particular issue is a big challenge. In this case, the question is finding whether the data is making assumption which makes it biased and renders it useless.

F. **Personally Identifiable Information**: Can we identify/extract enough information to solve a person's problem without violating his/her privacy?

G. **Challenges in Processing the data**: Collecting and aligning the data from various sources, transforming the data into a format which is suitable for its analysis, and understanding the output so as to derive some profitable knowledge from the data set are all problems associated while processing the data.

H. **Management Challenges**: Data Security and privacy; governance etc. are the major challenges faced while managing the data. The legal issues that Big Data providers face while storing/ processing/ accessing personal information of customers must also be considered.

I. **Data Sharing Challenges**: Speed limitations of transporting data though the Internet is a major challenge faced by Big Data Applications. Datacenters have the ability to store and compute large volumes of data, but bandwidth limitations during communication reduces efficiency during communication. This problem with data transmission speeds does not look like being solved any time soon. So, an alternative measure might be to reduce the data sizes while sharing it.

4. DATA SECURITY USING CRYPTOGRAPHY

Information has become more vulnerable due to the expansion of worldwide communication and digitalization. Due to this, sensitive information may fall into the wrong hands, and unauthorized personnel may modify data. If sensitive data falls into the wrong hands, the impact of the abuse can be much worse than in a paper-based system.The process of application of a 'transformation'/ modification to existing data in such a way that the data is readable only by the sender and the intended receiver is called **cryptography**. The transformation that is applied to the data is called as the **encryption algorithm** and the result of the application is called as the **ciphertext.** When the encryption key is the same as the decryption key, the algorithm used in the procedure are called as **symmetric ciphers** or **conventional ciphers**. When they are not the same, the algorithm is called as **asymmetric ciphers** or **public-key ciphers**. The public-key cipher is called so because, even if the encryption key is given to the public, the decryption key cannot be found and the data would still be undecipherable. Using public-key cryptographic methods, many algorithms have been devised to encrypt data. Secure storage and lines of communication have been developed to improve the security issues faced while storing and processing big data.

5. CLOUD COMPUTING

The model in which users' access services based on their requirements without regard to where the services are hostel is called as **utility computing** or **cloud computing**. [2] This type of computing allows users to buy/ rent resources like infrastructure, server hosting environments and other services on pay-on-usage basis.

Through cloud computing, each and every IT service is turned into a utility. The concept of cloud computing virtualizes resources and provides the service through the internet. Infrastructure as a Service (IaaS) provides a virtual infrastructure platform for the user to build his application on. Platform as a Service (PaaS) delivers a platform to compute on and facilitates application deployment on the cloud. Here, the user does not have to manage the underlying hardware and software layers. Software as a Service (SaaS) is a cloud computing solution which provides users with a finished software that is ready-to-use. It runs on the internet, so the need to download, install and run the software locally is eliminated. Even though documents stored locally on a computer system are accessible within the network, access outside the network is prohibited. This is where cloud computing comes into the picture. Using cloud computing, all the data and software programs are stored on remote servers and are available to the user from anyplace where there is an internet connection. Even if a person's personal computer crashes, the data that he has been working on is stored remotely and available via other computers. Another advantage of using cloud computing is that, any person who has permission, can view documents/ files/ data uploaded by others and sometimes even modify and make changes to them. Cloud computing reduces cost significantly as customers use resources on a subscription basis.

6. SECURITY PROBLEMS IN THE CLOUD

There are many security issues surrounding cloud computing environment including confidentiality, multiple security policies, dynamic of services, trust amongst entities and dynamically building trust domains. [4]

1) **Legal and Organizational Policies:** These refer to the requirements that specify what information is accessible to who and who can manage the content that is available. This can be a legal security policy and can govern the way users, security personnel and administrators behave. These policies should define and authorize security managers to monitor, analyze and investigate the computer systems available in the organization. [5] Also, these policies must define the consequences of violation of set rules and regulations so as to reduce risk.
2) **Physical Security:** It is very important for a cloud service provider to have a strong physical security in place. The hardware components may be in danger due to attack by people or by natural disasters. This security layer must be in place definitely regardless of the software security that being provided to customers. Physical security is provided usually though data backups, firewall systems, round-the-clock security.
3) **Data Security, Loss & Leakage:** Storage and management discrepancies may result in data leakage and may lack the necessary data destruction policy. Data leakage is prevented and storage is securely by providing various privileges to employees and customers. The customer organization must have full access and control permissions to restrict third-party access to their data that is stored online. Data must be secured by using the **Secure Socket Layer** (SSL), **Point to Point Tunneling** (PPTP), **Virtual Private Networks** (VPN).
4) **Technical Flaws and Technical Attacks:** Common configuration errors can cause a huge impact on security of content on the cloud because most of the content stored share the same cloud configurations. These flaws must be rectified as soon as they are encountered to prevent leakage and loss of data to third-parties and put it at risk. The signing-in mechanism (authentication) should be made secure as well. Multiple levels/methods of authentication would improve security of data. Sometimes, single point of user authentication is used. This method allows users to log onto their cloud service through their web application and prevents others from using their services. Hackers keep developing new ways to bypass security measures enforced by service providers. They develop hacking algorithms way faster than the service providers develop security algorithms. Cloud service providers usually do not disclose details of the location in which data is stored to their customers. This is a major concern as far as customers are concerned.

7. DATA CONFIDENTIALITY ON THE CLOUD

The following are some of the encryption schemes that have been proposed for improving data security:

1) Advanced Encryption Standard

Advanced Encryption Standard (AES) is an example of Symmetric Key Cryptographic algorithm that converts the available data files from the existing plaintext format to an incomprehensible format which is called as ciphertext (cannot be read by humans to prevent the unauthorized users from gaining access to the data files.)A data block which is 128 bits in size is converted into a ciphertext which is also 128-bits large. The key which is used for encryption is used for decryption as well. The size of the key varies from case-to-case. The four main operations of AES are:

i) **SubByte**: Each data byte is scrambled.
ii) **ShiftRows**: Each data row is scrambled.
iii) **MixColumns**: Each data column is scrambled.
iv) **AddRoundKey**: The encryption is done using the key.

Encryption is usually done before/in the Secure Socket Layer (SSL) for the transfer and sharing part. The owner of the data encrypts it with the symmetric key and then, uploads the encrypted data onto the cloud using Secure Socket Layer. The data is then stored/transferred by the cloud service provider. Retrieval of data involves authentication processes so that only the data owner/ someone who has rights to access the data can retrieve it. During the retrieval process, the encrypted data is passed through the decryption algorithm that uses the same symmetric key as used while encryption.

Users usually register with the owner of the data so that they gain the ability to sort and receive the data files through the cloud. Validity of user accounts is ensured through login processes. After the user is authenticated, the user is given a decryption key and another extra secret key. At this point of time, the owner of the data encrypts the data iteratively (done based on the encryption key). SSL is then used for protection while transferring all the data to the cloud. After this is done, the user is in a position to access the data. He/she is authenticated using a username-password scheme. After this, a request is sent to the cloud server. The server now verifies the details entered by the user and starts the process of retrieving the data. The decryption process happens in the same way the encryption happened, and the decrypted data is made available to the user. Since, during the process, the data is always in an encrypted format, the cloud service provider has little/ no knowledge of the key and therefore, data confidentiality is guaranteed. This method of encryption is pretty secure, but is very inefficient as compared to other encryption methods.

2) **Homomorphic Encryption**
In this form of encryption, the data files in the plaintext format are manipulated using algebraic techniques like multiplication and addition. Some of the homomorphic encryption schemes allow only one of multiplication or addition to happen. These schemes are called half-homomorphic encryptions (e.g. RSA and Paillier). Others allow both addition and multiplication simultaneously. We must keep in mind that operations that are usually performed on database relations do not usually work with encrypted data. This encryption scheme represents a relation with a public-key. An extra column is added to the relation which takes either 0 or 1 to be its data values. This extra column indicates the presence of the row in the database relation. Along with this, an algorithm is provided to the operators so that they can work on the encrypted data. This is done so as to enable cloud service providers to gather the required data even when it is encrypted. Since the output at the client's end can be extremely large, the client is asked to specify the number of rows that he/she wishes to see. This ensures that the output sent to the client has a high possibility of being in the result of the query.

3) Attribute Based Encryption
Attribute Based Encryption (ABE) assigns roles to all the parties involved in cloud computing. The three main roles assigned to parties are: authority, data owner and consumers.

i) **Authority**: The authority generates the keys used by the data owners to encrypt the data and for the consumers to decrypt the data. Certain attributes in the data files are used to generate the keys.

ii) **Data Owner**: The data owner, based on the key generated by the authority, encrypts the data

iii) **Consumers**: The consumers, based on the decryption key generated by the authority, decrypts the data.

The attributes of the user trying to decrypt a file are checked and matched with the attributes in the ciphertext, and if they do not match, the user is not allowed to decrypt the file even if he/she has the appropriate key. [5] This method of data encryption works even when the data shared on the cloud is owned by multiple users. This is achieved through a number of different methods. The most common method is one in which, a single user uploads the entire content on the cloud and provides the other owners with the decryption keys. Another method to achieve group sharing is by using a Group Signature. Any member of the group can sign messages while keeping their identity secret. Only the Group Manager can reveal the identity.

Dynamic Broadcast Encryption (DBE) enables the sender to encrypt the data and send it through a secure connection and only a set of users who are authorized can access this data.

The Cloud Service Provider is generally not trusted because he is not part of the users' trusted community. The group manager, who is a fully trusted entity, manages registrations and revealing the identity of the owner of the data. Group members refers to the people who are registered to the group under consideration. Dynamic Broadcast and Group Signatures are used together to achieve secure data sharing.

4) Proxy Re-Encryption
In this method, ciphertext that is in encrypted form is re-encrypted by another user so that the third user is in a position to decrypt it. This method is usually used along with Attribute Based Encryption Schemes or Homomorphic Encryption. The data owner generates two key pairs based on the underlying encryption principle (which can be homomorphic or ABE). Along with these two keys, a proxy re-encryption key is also generated for each user. After this is done, the data owner encrypts all the data files and also generates tokens that are used for authorization for each file. The file is then uploaded onto the cloud. When an access request is encountered, the cloud service provider checks the authorization credentials of the user. If the user is authorized, the data owner allows access to the data

During the access procedure, users send a request to a primary cloud where some verification processes take place. After these verifications take place, a request is sent by the first cloud to the second cloud which in turn will perform the encryption operations and send the results to the user.

5) Hierarchical Identity Based Encryption
This type of encryption is used so as to restrict authorized or partially authorized users from sharing their key with unauthorized users (to ensure data security). The five main steps that are followed in Hierarchical Identity Based Encryption are:

i) **Setup**: The master security code and public parameters are generated.
ii) **Encrypt**: The encrypted content is generated using the public parameters (that were generated in the previous step) and the identity set.
iii) **KeyGen**: Generates the secret key.
iv) **Decrypt**: Converts the ciphertext file back to its original form.
v) **Delegate**: The secret key is made available to the end use.

8. CONCLUSION

Data is expanding in volume, velocity and veracity at an alarming rate. Big data computation is one of the biggest innovations in the past decade. The potential of analyzed big data is endless and can be used in all walks of life. Cloud platforms act as a ladder to bridge the gap between data generation and data storage/ analysis. On-demand services and platforms are made available through the cloud and infrastructures can be provided as a service based on public need. Cloud computing, by itself, faces a number of security issues (like legal policies, physical attacks, data leakage etc.). Cloud computing platforms are now being used to store Big Data in many organizations. We looked at the different methods of encryption that are being used to secure data that is stored on the cloud. Even though many encryption schemes have been proposed, each of these schemes have their own deficiencies and weaknesses. In data sharing, security is the most important factor. This is because most of the data that is being shared via the cloud is sensitive in nature and illegal access to cloud data might be very harmful to the owner of the data. We found that Attribute Based Encryption and a Fully Homomorphic Encryption pattern are the best schemes available currently to ensure cloud data security. A neural data model was discussed which runs on the principle of fragmentation and encryption of sensitive data. A brief conceptual explanation about Hadoop was given. Hadoop provides a cost-efficient and flexible platform for long-term data and organizational growth. SecureSafe, a cloud-based data storage system was analyzed and its architecture, strengths and shortcomings were discussed.

REFERENCES

[1] Melanie Swan, “Philosophy of Big Data- Expanding the Human- Data Relation with Big Data Science Services”, *Big Data Computing Service and Applications Conference, IEEE 2015*.

[2] Rajkumar Buyya, Christian Vecchiola, S.Thamarai Selvi, “Mastering Cloud Computing- Foundations and Applications Programming”, University of Melbourne, IBM Reasearch, Australia Madras Institute of Technology, Chennai, India.

[3] Vinay Kumar Jain, Shishir Kumar, “Big Data Analytic Using Cloud Computing”, *Second International Conference on Advances in Computing and Communication Engineering, 2015*.

[4] Philogene A. Boampong, Luay A. Wahsheh, “Different Facets of Security in the Cloud”, *Submitted to San Jose State University andNorfolk State University*.

[5] Hussain Aljafer, Zaki Malik, Mohammed Alodib and Abdelmounaam Rezgui, “An Experimental Evaluation of Data Confidentiality Measures on the Cloud”, *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems- MEDES 2014.*