ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

Review On Big Data Storage Using Cloud, Security Issues and Threats and Corresponding Prevention Algorithms Related To Cloud Storage

[1] Nita Radhakrishnan [2] Mehul Awasthi
[1][2] Department of Computer Science and Engineering,
SRM University
[1] nitanitin@yahoo.com [2] mehulawasthi7@gmail.com

Abstract: **-- The dawn of the internet age has brought with it an increasing demand for data storage and management along with it, a host of new challenges. One of the major such challenges are to keep such a large amount of sensitive data secure. Some of these issues have been addressed by devising the concept of big data and cloud computing. This paper reviews ways to implement big data storage using the services provided by cloud computing, the security threats faced when doing the same and certain threat prevention algorithms that have been cited in earlier publications.**

Index terms: **- Big Data, Cloud Computing, Security, Threats**

I. INTRODUCTION

The 21st century has seen a rapid growth in various sectors such as healthcare, public sector administration, entertainment, retail, global manufacturing, and personal location data. Earlier storage facilities were required on a meagre scale and emphasis on security factors was not a top priority. But with the dawn of the internet age, demands increased as huge amounts of data began to be generated at a massive rate. Ninety percent of the data was generated in the past decade alone and it became impractical for general information technologies to store and manage such large amount of data because it exceeded the capabilities of these systems. Storage was implemented on a broader scale and hence it became pertinent to find ways to manage, analyze and process such large amounts of data and also provide better measures to keep the data secure in order to quench the ever growing demands. The existing database storage systems could be scaled up or down but at a large cost. This scaling of the database storage systems could not be done without the commodity hardware in parallel. This was impractical because such hardware could not deal with the constantly growing data volume. To tackle these issues the concept of cloud computing was instrumented. It has met criteria such as ability to scale up or down, flexibility and cost effectiveness.

II. STORAGE OF BIG DATA USING SERVICES PROVIDED BY THE CLOUD

There has been an immense increase in the data we generate over the past decade. According to some estimates about 90% of the world's data was generated in the past two years alone[1]. The three big sources of big data are scientific data, networking data and business data[2]. For example the Internet which is one of the biggest sources of networking data generates a large amount of data, at an incredible rate. Facebook alone generates petabytes of data in a single day [2]. This obviously demands advanced technologies which will be able to manage such an amount of data in terms of acquiring it, storing and processing it. Most of the big data generated is in the form of unstructured data(about 80%). Businesses can gain valuable insight by analyzing such large amounts of data which in turn can help them make more informed decisions. For example, a business can analyze market reaction to a product it has recently launched by studying how the consumers' feedback to it on social media via tweets, Facebook posts etc. Depending on the ways consumers react, it would obviously help shape their future decisions and marketing policies[3]. As can be seen, analysis of data has a huge potential. Not just for businesses but for healthcare, scientific study, government sector, e-commerce and so on and so forth.

Now come the four unique problems posed by big data, also known as the 'four V's'. Volume, velocity, veracity and variety. Volume refers to the large amount of data produced. Every day the amount of data produced increases exponentially. It ranges from petabytes to Exabyte. Velocity

ISSN (Online) 2394-2320

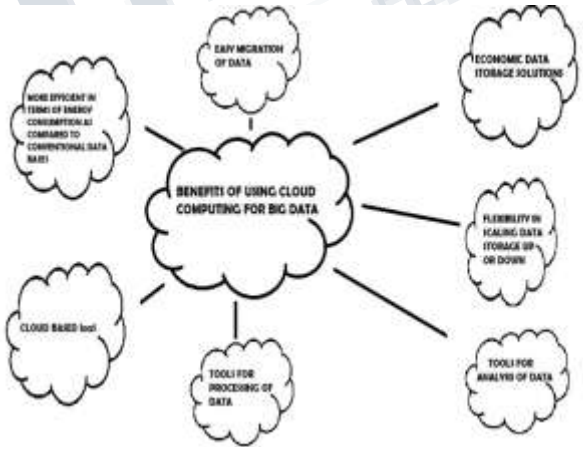
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

is the rate at which data is generated. Data is produced at a very high rate and sometimes there also a necessity to process it quickly. For example when trying to stop a fraud from occurring, it is important that data be processed as fast as possible. Veracity is the credibility of the source of data. If the source of big data is not trusted, then decisions based on the data cannot be taken. Trustworthiness in big data is another challenge with the ever growing sources [3]. Variety refers to the different types of big data generated, which can be structured, semi structured, unstructured and also variations in size and formats.

Big data poses a new set of challenges. Conventional storage systems cannot cope up with the storage or processing demands of big data. The existing data base management systems can be scaled up with additional hardware; however this is a very expensive and uneconomical process, both from a financial and environmental perspective. However with the recent developments in advanced data base technology, wireless data and mobility it is now possible to address these challenges posed by big data. Cloud computing can be deployed to serve as a solution to these problems and provide the necessary infrastructure.

Cloud computing can provide cheap storage solutions for storing the data, along with a host of other services and tools for the analysis and processing of the data. Because of the affordability and flexibility provided by cloud computing many businesses are attracted to it. Such advancements in technology have compelled them to move to new service paradigms such as Software-as-a-Service, Infrastructure-as-a-Service etc. Currently some cloud based Iaas (Infrastructure-as-a-Service) are being used by companies to tackle the big data problems [1]. Examples of these are Google Big Query, Rack space Big Data Cloud and Amazon Kinesis. Google Big Query is used to compute, analyze and store big data. It is very powerful and provides real time insights of business data and is capable of extracting information in a matter of seconds from multi-terabytes of data [1]. These services are hence useful for companies to store their structured and unstructured data cheaply and analyze it at the same time, as they provide both storage and computational services. This is also of immense advantage to the organizations which are testing and developing their applications and systems. Cloud computing therefore plays a huge role in the setting out of big data [3].

The benefits of big data in cloud computing are aplenty. Cost effectiveness is the most obvious benefit. Organizations do not need to spend funds on scaling their existing infrastructure up to manage the growing amount of data, nor do they need to spend money on the maintenance of such infrastructure. Cloud computing provides a cheaper yet much more efficient alternative. Investment on such cloud paradigms would be significantly lesser. It provides resources on demand with cost according to usage [3]. Cloud based data systems are also more efficient in terms of energy consumption as compared to conventional data bases, hence making them economical and environmentally friendly. The other major distinction of using cloud computing would be the flexibility it provides. Storage can be scaled up or down rapidly, depending on the organization's need, without much of a fuss. This would be unthinkable and a very expensive process for a conventional data base system. Data can also be migrated easily.



III. SECURITY ISSUES AND THREATS FACED WHILE USING CLOUD FOR STORAGE

Cloud computing is predicted to transform the computing world from using local applications and storage into centralized services provided by various organizations. However with storage of such massive data on a centralized basis raises questions regarding the safety, privacy and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

confidentiality of the data stored. To fully understand the answers to the posed questions we must first realize the different security issues faced in cloud computing. In general the security issues faced can be divided into two categories, the ones faced by the cloud service provider (CSP) and the ones faced by the users of the cloud. Some of the issues are as stated below [4].

A. Trust

This refers to the mutual trust between the user and the cloud service provider. This is clearly demonstrated by the following example. Consider two entities A and B where A trusts B and hence expects B to deliver valid and reliable services as requested by A. Here let A and B be the user and the Service Provider respectively. Then the user's confidence depends on the integrity of the service and the efficiency of the security measures. Also working as per the rules and protocols and providing the user with a caution that a minimum risk may occur at any time is compulsory.

B. Confidentiality and privacy

Privacy is an essential right that has to be exercised as a world that's devoid of privacy will result in loss of data and will impose a stringent limit on the freedom that one possesses. Confidentiality is defined as the ability for an authorized group of users to access protected data.

C. Integrity

Integrity is of two types with reference to data stored on the cloud. They are as follows

- ***Data Integrity:*** it refers to protecting the data from unauthorized access and modification.
- ***Software Integrity***: It refers to protecting the software from unauthorized access and modification.

D. Availability

This refers to the availability of data and other resources to the authorized users Some of the security threats posed by malicious attackers as mentioned in the paper are as follows.

- The CSPs are abused and corrupted in order to infiltrate or corrupt the user's property and this is done by hosting malicious code into the CSP. These types of attacks are commonly used on Paas and recently Iaas.
- Corrupted or malfunctioning APIs are also a source of hosting malicious code with the intention to corrupt user property.
- More often than not, CSPs face threats from inside attackers such as employees of that particular organization who have access to the data and decide to misuse their position and power.
- Some resources such as those provided by Iaas are shared between users. For example, multiple operating systems running on the same system, can intentionally or unintentionally cause damage to the existing system properties.
- Data loss or leakage is another major factor to be considered. This can happen if the data is unintentionally or intentionally deleted without appropriate backup.
- Corruption can happen if malicious attackers gain access to user accounts or profiles by phishing, fraud, and exploitation.

IV. REVIEW OF CLOUD SECURITY ALGORITHMS

The need for flawless mechanisms to secure the data being stored in the cloud has become imperative owing to the issues discussed in the previous sections. The challenge lies in developing security algorithms with a strong foundation. Encryption is the process of converting plaintext message into an encrypted text called the cipher text and an encryption algorithms along with the key is used to obtain encrypt and decrypt the data. Encryption designs are founded on block or stream ciphers [5]. There are many such algorithms, namely, DES algorithm, RSA algorithm, AES algorithm and MD5 algorithm. This section reviews each of these algorithms which $_{h0ave}$ been proposed as solution to above stated security issues.

A. DES Algorithm

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

The data encryption standard was developed by IBM in 1974. It has a 64-bit block size key out of which only 48 bits are used since 8 bits are control bits, and is a symmetric cryptosystem that follows 16-round Feistel Network for Round Computation. According to this algorithm both the sender and the receiver must be aware of the same encryption key. In order to implement this algorithm, we will need the round function, the key schedule and the initial and final permutation rounds. The initial round permutes the accepted 64-bit plaintext (Px) and changes the bit positions. The final permutation is the reverse of the initial permutation. It maps the bits in the opposite manner to which the initial round permuted it.Round function using Feistel Network splits the Px into two parts, namely L_0 and R_0, each 32 bit long. The expansion box is applied to R_0 after which it is XORed with 48 bit keyword. This 48 bit keyword is generated from the 56 bit keyword from a series of left shifts followed by compression boxes.

It complies with the properties of every block cipher, i.e. a small change in the plaintext affects the cipher text greatly and every bit of the cipher text depends on many of the previous bits of the plaintext. Therefore for data transmission over a cloud network or any network such as phone lines that has some form of disturbance, the error or disturbance will be propagated to all the subsequent blocks as each block is dependent on the previous block. But this mode of securing the data is more effective compared to the Electronic Block Code (ECB) since this has an extra XOR step that adds one more layer to the encryption process.However recent developments of various other algorithms have proven to be more efficient, owing to the exhaustive key search due to its key size and its inability to prevent brute force attacks, differential cryptanalysis, linear cryptanalysis and Davie's attacks.

B. RSA Algorithm

RSA algorithm is a widely used Public-Key Algorithm and was first publicly described in 1977 by Ranold Fivest, Adi Shamir and Leonard Adleman. Using this algorithm, we can encrypt the data in order to prevent unauthorized access to confidential information. RSA is a block cipher in which every message is mapped to an integer. Let the decryption key be DEC and the encryption key be ENC and the message to be transmitted be msg and the corresponding cipher text generated from the encryption be cipher, then

DEC[ENC[msg]] = msg
ENC[DEC[msg]] = msg
ENC -> publicly available to all
DEC -> it is private to only the receiver
Cipher = ENC[msg]

The encryption key ENC is stored in a public file available for sharing between various users. The DEC key is private to every individual and hence can be accessed only by the intended receiver. This thereby avoids the problem of eavesdropping.

The procedure for generation of ENC and DEC is as follows,

- Assume two large prime number p and q and multiple them to obtain n=pq
- Now we find the totient of n which is calculated as $\varphi(n) = (p-1)(q-1)$
- Find a number e such that $1<e<\varphi(n)$ and e is co-prime to $\varphi(n)$.
- Now we compute the modular multiplicative inverse of $e \pmod{\varphi(n)}$ as d.
- Therefore ENC = msg^e and DEC = $cipher^d$

It has been proven that the procedure to obtain the decryption key is extremely difficult unless some prior information is available. However it can be vulnerable to attacks if the two prime number chosen are small. Also RSA is a deterministic algorithm and does not have any random component. This can also prove to be a vulnerability. But the recent developments show that the random padding scheme resolves these issues making this algorithm quite popular.

C. AES Algorithm

This is a symmetric key encryption standard. It is also called Rijindael [6].Each block of the cipher is 128 bit long, and the key sizes can be 128, 192 and 256 bits, respectively. All rounds of AES are similar except the last one. The AES works on 4X4 matrices and it consists of key expansion, initial and final round. Add Round Key, Sub bytes, Shift Rows, Mix columns form a part of the initial while the final round is similar to the initial round except that

it does not have the Mix columns. AES works well on both software and hardware. This algorithm is vulnerable to security attacks as cloud computing systems are providing a wide variety of services to enable vendors to rent out spaces on their physical machines. The imminent threats include, using **DDoS** attack (many node systems attacking one node all at the same time with a flood of messages) to exhaust server resources, using known vulnerabilities to interrupt services

D. MD5 Algorithm

This is a message-digest algorithm as it transforms an input of arbitrary length into an output of constant length, and is a widely used cryptographic hash function. It was developed in 1991. It takes in input of arbitrary length and uses 128 bit hash value and is expressed in text format as 32 digit hexadecimal number. The usual length of the output produced by this algorithm is 128 bits. It follows 5 steps.

- ♣ It appends the padding bits
- ♣ It then appends the length
- ♣ It initializes the MD buffer
- ♣ Process message in 16 word blocks
- ♣ It finally produces the output

This algorithm is commonly implemented in some UNIX based system as utility md5, class MD5CryptoServiceProvider in Microsoft's.NET framework, etc. MD5 is also used in conjunction with other cryptographic methods in digital signature applications or in protocols like SSL and others.

REFERENCES

[1] Muhammad Adnan, Muhammad Afzal, Muhammad Aslam, Roohi Jan, Martinez Enriquez A.M, "Minimizing Big Data Problems using Cloud Computing Based on Hadoop Architecture", High-capacity Optical Networks and emerging/Enabling Technologies (HONET), 2014, 11th Annual, December 2014, Page(s):1-2

[2] Han Hu, Yonggang Wen, Tat-Seng Chua, Xuelong Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", IEEE Access Vol. 2, June 2014, Page(s):1-2

[3] Vinay Kumar Jain, Shishir Kumar, "Big Data Analytics using Cloud Computing", 2015 Second International Conference on Advances in Computing and Communication Engineering, May 2015, Page(s): 1-5

[4] Louai A. Maghrabi, "The threats of data security over the cloud as perceived by experts and university students", page(s)-2-4

[5] Ms. Theres Bemila, Karan Kundar, Lokesh Jain, Shashikant Sharma, Nayan Makasare, "Comparitive study of security algorithms in cloud computing", ISSN (online), vol. 5, issue 3, march 2016

[6] Er. Ashima Pansotra and Er. Simar Preet Singh, "Cloud Security Algorithms", International Journal of Security and its applications, Vol.9, No.10 (2015), pp. 353-360