ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 1, January 2017

Big Data Security

[1] Sheryl Saji [2] Ann Sara Sajee [3] Akshada Potdar
Department of Computer Engineering,
Fr.Conceicao Rodrigues College of Engineering,Vashi,Navi Mumbai

Abstract: **-- Big data is a collection of relevant data that happen to be very important for any organization .If these data are changed or tampered illegally it can create great losses .It is very important to store the data properly. There are various attacks planned on them which if successful can be very profitable for the attackers. In this paper we have discussed about the importance of data security and also its practical application. We have also discussed the various tools that can be used to secure the data. Thus keeping the customers data safe.**

I. INTRODUCTION

Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Big data terms describes the large amount of data in both structured and unstructured form. Big data includes various challenges such as analysis, capture, search, storage, visualization, and privacy violations. Big data handling is moreover difficult using the relational database management system. Big data security is an major concern when we deal with a massive data. Big Data can be described by using the following characteristics

1. Volume

Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-tomachine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden. The volume of data, especially machine-generated data, is exploding, how fast that data is growing every year, with new sources of data that are emerging.

2. Velocity

Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

3. Variety

Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. This also help people for deeply analyzing the data associated with it and make use of various advantages related to it. A lot of this data is unstructured, or has a complex structure that's hard to represent in rows and columns. And organizations want to be able to combine all this data and analyze it together in new ways.

4. . Variability

In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. This can be difficulty for the people who does the analyzing of data.

5. Complexity

Managing of data is difficult when large amount of data is consider from multiple sources. The data from the multiple sources is the reason which leads to the difficulty in linking, matching and transforming data across the various systems. Hence, it is a need to connect and correlate relationships and multiple data linkages Applications:.Healthcare[2]: It helps to refer through the old treatment data for similar symptoms and diagnose them faster and more accurately Retail: It is a predicting trends, for forecasting where the demand will be for those products, optimizing pricing Banking:. Big data analytics is being successfully used in banking sector with the various aspects such as spending pattern of customers, sentiment and feedback analysis, security and fraud management ,etc.

II. WHAT IS BIG DATA SECURITY

A significant portion of information security efforts go into monitoring and analysing data about events on servers, networks and other devices. Big data security analytics is qualitatively different from other forms of security analytics. The need for tools for integrating and visualizing diverse types of data,. Large

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

amount of data from various application such as healthcare, retail, and banking along with the personal data is transferred over the network. So naturally the concern about privacy and security arises Data cannot be sent encrypted by the users if the cloud needs to perform operations over the data. The solution used for this is "Fully Homomorphic Encryption" which allows data stored in the cloud to perform operations over the encrypted data so that new encrypted data will be created.

Different security and privacy challenges are :

1. Secure Data Acquisition

Data acquisition is the process of acquiring the data. The data obtained from source will have all the characteristics of big data- volume, velocity, variety etc. Hence, it is difficult to guarantee the security of the data. Many times some malicious code may be present in some unstructured data which may be difficult to find out. Hence, getting a secure data at data acquisition step in itself is difficult.

2. Secure Computation:

While computing the large amount of data, the data is mostly partitioned. At such stage, distributed data are executed in parallel to complete the data computation. In a map reduce framework, the data is splitted into different files which are then computed in parallel using mapper. In this case, the data is distributed and hence, had to be secured from any data fragment being attacked by any external source.

3. Secure Data Storage:

Huge chunk of data is getting generated every minute!. In order to store the data itself is getting extremely difficult. Moreover if the security of the data s not high, it may result in attack on the confidentiality and integrity of the users data. Privacy challenge for big data is extremely high.

4. Secure Input Filtering:

Data comes from various sources so it is very important that the source is reliable or else it can lead to great losses.

5. Secure encryption of data:

Big data may contain some extremely sensitive information which has to be protected. Hence, it is important to apply some cryptographic algorithms which will encrypt the data so as to secure them. However, many traditional encryption algorithms like Blowfish, DES, Rijndael have significant bottleneck in the light of Big data. Since, big data has huge amount of data and real time data processing is required, there is need for an encryption algorithm which works faster and is effective.

6. Secure Access control:

The data present may have sensitive information which if accessed my any one will lead to challenges in privacy and security of data. Hence, these data requires access control from people who are only responsible to access the data.

7. Anonymity Concerns

Many customers may feel uncomfortable with the idea that businesses are able to collect such detailed information about their identities, behaviours, motivations, and other sensitive facts. Some companies respond to these concerns with data masking policies and aggregating data sets, but those methods aren't always effective so, the right tools are required.

8. Data Breaches Are Now Common:

Large amounts of data are a goldmine for cyber criminals, and companies that collect and store it are big targets. The recent Sony hack is another example, with experts estimating as much as A00 terabytes of data stolen or leaked. Data breaches are now far too common and will likely not go away anytime soon.

9. Data Brokers

Even if a company goes to great lengths to protect big data, if they sell some of that data to third parties, risks could increase. Currently there are few laws that address brokered data, which certainly compounds the problem. There are also few ways to make sure data brokers are held accountable for the data exchanged between parties, making it an even more challenging task to ensure all big data is protected. With enough resources, skilled workers, a detailed big data strategy, and a commitment to customer privacy, many of these concerns can be directly addressed. If these problems are solved, businesses will be in a better position to truly take advantage of all that big data has to offer.[5]

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

III. IMPORTANCE OF DATA SECURITY

Data security is a very important issue for businesses and organizations today. If some unauthorized user gains access to the data of an organization, then the organization may suffer serious problems. The sources for security breaches have increased exponentially over the past few years. For example An authorized user may use the credit card number of another user for shopping. He can also delete important data of a business or an organization. .This can be done by: Building and maintaining a secure data network Regularly monitor and test networks and Maintain an Information Security Policy[3]

IV. TOOLS FOR DATA SECURITY

The challenge of big data encryption is that, while there are plenty of encryption offerings around, but these offerings don't secure all the log files and configuration information associated with the big data environment itself. These approaches also tend to introduce a significant performance hit.. The other challenge is that after sometime we require to dispose the old data, as it can link to the individual's current details.

A. HADOOP:

Big data that resides within a Hadoop environment can contain sensitive financial data in the form of credit card and bank account number proprietary corporate information and personally identifiable information (PII) such as the names, addresses and social security numbers of clients, customers and employees. As such, sensitive data was safely confined in isolated clusters or data silos where security wasn't an issue. But that quickly changed as Hadoop evolved into Big Data as-a-Service (BDaaS), took to the cloud, and became surrounded by an evergrowing ecosystem of tools and applications that poses new security challenges like - Ensuring the proper authentication of users who access Hadoop. Ensuring that authorized Hadoop users can only access the data that they are entitled to access.. They get data access histories for all users are recorded in accordance with compliance regulations and for other important purposes.. Ensuring the protection of data—both at rest and in transit—through enterprise-grade encryption.

Hadoop Security Best Practices:

1. Plan before you deploy

Big data protection strategies must be determined during the planning phase of the Hadoop deployment. Before moving any data into Hadoop it's critical to identify any sensitive data elements, along with where those elements will reside in the system. In addition, all company privacy policies and pertinent industry and governmental regulations must be taken into consideration during the planning phase in order to better identify and mitigate risk

2. . Don't overlook basic security measures

Basic security measures can go a long way in meeting Hadoop security challenges. To ensure user identification and control user access to sensitive data it's important to create users and groups and then map users to groups. Permissions should be assigned and locked down by groups, and the use of strong passwords should be strictly enforced. Fine grained permissions should be assigned on a need-to-know basis only and broad stroke permissions should be avoided as much as possible.

3. Choose the right remediation technique –

When business analytic needs require access to real data, as opposed to data that has been desensitized, there are two remediation techniques to choose from—encryption or masking. While masking offers the most secure remediation, encryption might be a better choice as it offers greater flexibility to meet evolving needs. Either way it's important to ensure that the data protection solutions being considered are capable of supporting both remediation techniques. That way, both masked and unmasked versions of sensitive data can be kept in separate Hadoop directories if desired.

4. Ensure that encryption integrates with access control

Once an encryption solution is chosen it must be made compatible with the organization's access control technology. Otherwise, users with different credentials won't have the appropriate, selective access to sensitive data in the Hadoop environment that they require.

5. Monitor, detect and resolve issues

Even the best security models will be found wanting without the capability to detect noncompliance

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

issues and suspected or actual security breaches and quickly resolve them. Organizations need to make sure that best practice monitoring, and detection processes are in place.

6. Ensure proper training and enforcement

To be fully effective, best practice policies and procedures with respect to data security in Hadoop must be frequently revisited in employee trainings and constantly supervised and enforced.Hadoop is enabling organizations to analyze vast and rich data stores and derive actionable insights that inform new and better products and services and help to create competitive advantage. But the benefits of Hadoop come with risks. Hopefully the above information will help organizations to gain a better understanding of the security and compliance issues associated with Hadoop and to implement best practices to keep sensitive data safe and secure going forward.[4]

B. VORMETRIC :

To provide big data analytics security for these confidential assets, security teams can use the following solutions: Vormetric Transparent Encryption can easily be deployed on servers, where it can encrypt big data outputs and control and monitor who accesses them. You can use the Vormetric Application Encryption to secure specific fields that may be created in analytics applications. Big data frameworks. Within the big data environment itself—whether it's powered by Hadoop, Mongo DB, NoSQL, Teradata, or another system—massive amounts of sensitive data may be managed at any given time. [7]

C. NO SQL (NOT ONLY SQL) TECHNOLOGY

No SQL is another technology which is widely used to handle the big data mostly unstructured data. It basically provides flexibility and scalability. There is no schema used in No SQL databases which is very helpful for dealing with huge amount of unstructured data. Main databases under No

SQL technology:

1. Key value pair store
2. Document oriented databases
3. Graph databases
4. Extended Relational databases

1. Key Value Pair store:

It stores the Key value pair is used to store the huge unstructured amount of data. There is no scheme used in the key value pair store like in RDBMS which provide great flexibility. There is one unique key and a particular value corresponding to that. The key is an entity and value is attributed in key value. Key value pair is also used in map reduce by map function. After loading of data map function fetch data in key/value pair in Hadoop environment. Most data is stored in string in key value pair store.

Example: Key Value Facebookuser7778_name Jas Twitteruser4566_name RohitRiak is the key value pair database which implements the No SQL technology and it is widely used among the social networking websites to handle its data.

2. Document Oriented Database

Document oriented is one of the famous No SQL database. It uses the document to handle the data. Mongo dB is the open source famous No SQL database. Mongo dB is used JSON (Java script object notation) and BJSON (Binary JSON). JSON is used for writing the queries. It is used to handle a document. In Mongo dB data are stored in a document with no schema and also without the concept of normalization. In mongo dB tables are automatically created after inserting data into documents via JSON. Tables are known as collection in mongo db.

3. Graph Database Graph database

It is used to store the data in form of nodes and edges. It is a very easy and better approach than RDBMS due to its processing convince because graph database is very fast in terms of performance as compared to RDBMS. It is very helpful in performing graph like queries.

4. Extended Relational Database

Nowadays unstructured data is growing very fast. No SQL database, HDFS and other databases come into the picture to handle this data. Now to compete with these databases, companies like Microsoft and Oracle also extending their databases to remain in the market. Microsoft's file table is a good example of it. They are basically trying to make their software compatible with unstructured data [6].

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

D. CLOUD STORAGE SOLUTION FOR BIG DATA:

1. Cloud storage:

Cloud storage is considered as a method to store big data or huge volume of data. In today's age, everyone stores all their sensitive information on cloud. The data from users across the globe are stored in any part across the globe. Different countries have different privacy laws. So it remains unclear which laws of which country regulate that data privacy while it flows from the sender to the server. Users often feel their data is secure. However they store their data in a place which does not belong to them. So there may be some circumstances in which legal permissions may be given to access data on cloud.

2. Cloud Service Models:

There are different ways in which the cloud can be serviced to the users. These are – Software as a Service(SaaS), Platform as a Service(PaaS), Infrastructure as a Service(IaaS)

3. Data Security Challenges in Cloud:

People use the cloud for cost saving and new agile business models. However, the security of the data under many circumstances is compromised.[9] In the cloud computing environment, it becomes particularly serious because the data is located in different places in the globe. Data security and privacy protection are the two main factors of user's concerns when data is stored on cloud. There are complex data security challenges in the cloud:

4. Confidentiality of data:

Many business, government and regulatory data are present in the cloud which under any circumstance cannot be compromised. Sensitive government data if gets under wrong hands will cause huge loss to the country. Similarly for business as well. Hence, data had to be confidential to the owner of the data only. So, security of the data should be high.

5. Data Infrastructure sharing:

The cloud belongs to all and all can put their data on cloud. The multiple tenants sharing the same infrastructure will cause the cloud storage process complex.

6. Legislative rules and regulations:

The cloud server on which the data is stored is available under certain rules and regulations which under worst case may harm the confidentiality and integrity of the data.

7. Data Protection on Cloud:

Encryption along with other core data security technologies, increases security, provides a comprehensive multi-layered approach to protecting sensitive data and mitigate risk in and out of the cloud. Hence, data centric approach should encrypt data, have key management, strong access control and security to protect data on cloud. Avoid putting sensitive information on cloud. Data lock the sensitive information if on cloud. Read users agreement policies thoroughly to understand cloud service providers access policies. Security Intelligence. Use an encrypted cloud service.

V. ISSUES IN TECHNOLOGY:

An additional big data security challenge is that big data programming tools, including Hadoop and No SQL 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. No SQL databases lack some of the security features provided by traditional databases, such as rolebased access control. The advantage of No SQL 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.[1]

VI. CONCLUSION:

Data security is the need of the hour especially when the entire world is very dependent on data. The data we deal with is usually big and can be very hard to handle. Hence we use the various Data security tools that are currently available, we have discussed its advantages and how it can be used for various types of applications.

REFRENCES

[1] K Jayabharathi - ijrset.in

[2]https://www.healthcare.siemens.com/magazine/ mso-big-data-and-healthcare

[3]http://searchsecurity.techtarget.com/magazineContent/Big-data-security-analytics-Harnessing-newtools-for-better-security

[4]https://www.qubole.com/blog/big data/hadoopsecurity- issues/

[5] https://www.qubole.com/blog/big-data/bigdata-security-concerns/

[6] greatinformer.blogspot.com/20A2/AA/datasecurity-importance-of-data.html

[7]http://enterpriseencryption.vormetric.com/rs/vormetric/images/wp-cso-vormetric-data-security-inthe-cloud-updated.pdf