Big Data Privacy for End to End Delivery

Ashutosh Dixit, Nidhi Tyagi

Abstract-: Data privacy is an area of concern to process massive datasets in Big Data applications. Assortment of Big Data-sets is tough to be handled using, with the use of on-hand management tools or traditional processing techniques, the assortment of Big Data sets is difficult to be handled using Big Data has three characteristics i.e. V's Volume, Varity, and Velocity . Privacy to such Big Data could be a massive snag which might be achieved by Anonymization technique. Datasets like financial data, Health Records and other confidential information of various organizations needs privacy to protect from the intruders and malicious entities.

The aim of Big Data Anonymization is to shield the privacy of the individual and make it legal to share the information while not obtaining permission from people.

The research paper discusses the basics of Big Data, technology behind it and various challenges.

Keywords: Big Data, Hadoop, HDFS, Map Reduce, Data Anonymization, Kerberos Security System.

I. INTRODUCTION

Big Data is a self-explanatory word which itself means a collection of data that is large in volume, complex in structure and generated with tremendous speed. These data sets are too complex to be handled with the traditional or existing data systems. It can be examined for the solution that proceeds for better decisions and planned business moves, or to the business problems for which we are not capable to solve before.

Volume, Velocity and Variety represents the basic Features of Big Data [1] [2] [3].

Volume: It represents the amount of Data, which is increasing day by day. The sources of this data are m/c

Scientific instruments, social media etc. Researchers predict that 40-50 ZETTABYTES will be generated by 2020, which has increased from 2005 to 300 times.

Velocity: It represents the speed of data which can be find and acted on. Some internet based products needs real time evaluation and action. Velocity deals with the growth of data and the production of handling the real time data.

Variety: represents the type and nature which can be in structured and unstructured formats for example texts audio, video we have to need extra preprocessing for find out the meaning.

During the past few years, above mentioned three important Vs of Big Data two more Vs identified are veracity and value [6].

1.2 Challenges of Big Data

Provides big data a lot of promises and adds on the advantages for handling the complex data, but it is never possible for the Big Data to do it without facing the challenges. Some of the challenges are discussed below.

(i) Data Quality: -

The problem here that big data is dealing with is veracity as most of the data is unstructured and it is incomplete, inconsistence. In it, it can be complicated to search and analyze for documents and photo, audio, videos and other unstructured data.

(ii) Discovery:

Organizations can achieve its goal & sustain in the competition if they are capable to extract insights from their big data and then can work upon the quickly.

(iii) Storage:

Simply it is the big challenge for big data to store and analyze all the information. Huge amount of data in any organization more it become difficult in managing and storing the large volume of data.

(iv)Analytics: -

As most of the times data comes in an unstructured form and we are unaware of the nature, kind, type or format of data, it is quiet for analyzing the data.

(v)Security: -

For the big data storage in organization, security is the big concern. It includes identity and access control, restricting access based on a user, data encryption and data segregation.

(vii)Lack of Talent: -

The organizations can develop these big data applications. They can also manage and execute these applications which provide the insights. For this work the organizations have to appoint professionals and experts of big data skills. It creates the demand of big data experts which in result demanding big salaries.

II. HADOOP: SOLUTION FOR BIG DATA PROCESSING

It is an open source framework that provides Big Data and it should be stored and processed in the distributed environment throughout the use of simple programming models of computer groups. It has been created to scale thousands of single servers, offering each local calculation and storage. It lives in a local file system of the personal computer. In Hadoop, the data resembles and it presents in a distributed file system that is known as a Hadoop distributed file system [4].

Revised Version Manuscript Received on August 19, 2019.

Ashutosh Dixit, Research Scholar, Bhagwant University, Ajmer, Rajasthan, India.

Dr. Nidhi Tyagi, Professor, MIET, Meerut, U.P. India.

2.1 Hadoop Architecture

Hadoop framework includes:

(i) Hadoop Common

It represents to the collection of common utilities and libraries that is very supportive for other Hadoop modules. This mainly consists of java libraries, jars and other important utilities that is required by other modules of Hadoop architecture. Hadoop Common package is also used source code and for documentation. Hadoop common accepts that the hardware failures are common problems and it is the responsibilities of the Hadoop Framework to control the problems in software automatically.. Hadoop Common is also known as Hadoop Core.

(ii)Hadoop Yarn

It provides a framework for job scheduling and cluster resource management; it is the responsibility of YARN to provide the computational resources (e.g., CPUs, memory, etc.) required software executions. It resides between HDFS and the processing engines being used to run applications.

(iii) HDFS

Hadoop Distributed File System gives the facility of high throughput access to application data and it is used for applications that are in big datasets. It is depend on the Google File System (GFS). It states that file will be broken in to blocks and stored in nodes (NameNode and DataNode) over the distributed architecture which helps in providing high-performance access to data across highly scalable Hadoop clusters. It is specially designed to be highly fault-tolerant and also designed to be deployed on low-cost hardware [8].

(iv) Hadoop MapReduce

A framework or a processing technique which is used to process big data sets by writing java based programming model. The MapReduce algorithm includes two significant tasks i.e. Map and Reduce [5]. Figure 1 represents the components of Hadoop.

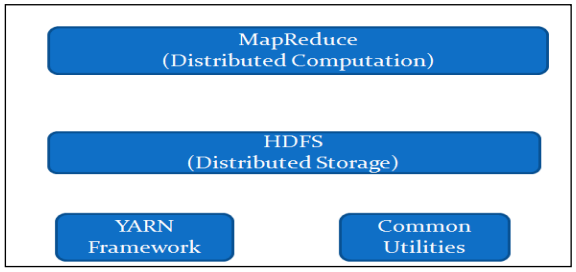

Figure1: Components of Hadoop

III. HADOOP SECURITY

(a) Security of Data during Motion

In Hadoop system , data node store the actual data in a distributed manner whereas the name node has meta-data and edit-log stored to keep the record of data nodes. The real communication of data blocks takes place between client node and Data node. So it can be said that Hadoop framework has several nodes with data communication between them. Therefore, the default data is not encrypted and the data when it is being transmitted at this time it is open and can easily to attack by hackers, trackers and crackers.

For internodes communication, various communication protocols like Remote Procedure call (RPC), Transmission control Protocol over Internet Protocol (TCP/IP), and Hypertext Transfer Protocol (HTTP) has been used. Some solutions are also available for securing communication various nodes like Kerberos and Simple Authentication and Security Layer (SASL) [11].

(b) Data Security at Rest

It means that data is being stored in certain storage. By default Hadoop cannot encrypt data that is stored on disk and that can show sensitive data for security attacks. This is really a major problem because of the nature of Hadoop architecture [7], which spreads data across a multiple number of nodes. The data blocks are exposing some unsecure entrance points.

(c) Loss of Control

Companies are unable to know if their data is being used by someone else because they are not capable to look and make their direct control over the data because of number of transparent mechanism to monitor the resource directly. It creates the various security problems after they discontinue the use of services their data is not fully removed by their service provider.

(d) Trust chain reliability in Clouds

Customers cannot make direct control on over their data because they have to share their physical cpmponents with other customers. For this reason they believe on the cloud providers. Trust mechanism is a substitute by providing to the user transparent control over their data and cloud resources.

(e) Privacy & Security Concerns in Clouds

Both are two different domains but they are related to each other for this reason they are discussed together. There are some needs of Security so it is responsibilities of the enterpriser and they are sure that their sensitive data is not being accessed by distributed environments. They also need to assure because of money that it is not being shared to some third party. It is the most important reason for which an organization or a company has to secure their data with Hadoop environment [10].

IV. AVAILABLE SOLUTIONS& RESULTS

There are many solutions of data security in big data of its feasibilities and obstacles. Besides it, we also discussed intelligent analytics to security concern issues in over security intelligence architecture.

4.1 Kerberos Security System

To serve it provides the service of strong authentication to client/Server application without transferring the Password over the network Kerberos. Kerberos is a secure and authentic network. It works by using time-sensitive tickets which are generated with the help of Symmetric Key Cryptography, represents in figure 2.

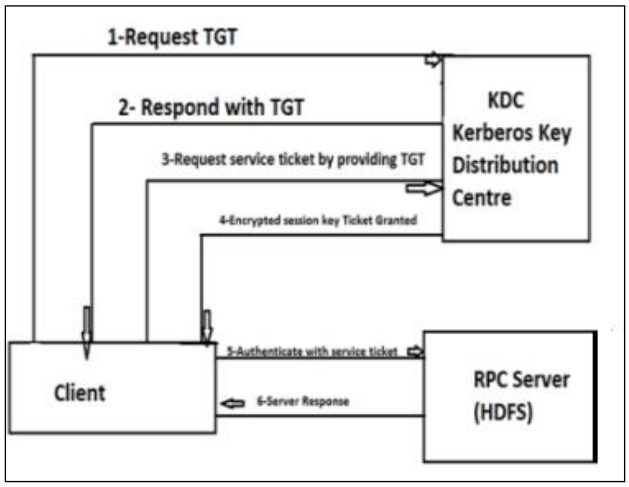

Figure 2: Kerberos System

The term "KERBEROS" has been taken from Greek mythology which means the three headed dogs, these dogs work as the guard on the gates of Hades [10]. In the some way Kerberos has three heads security paradigms as given below:

(i) The user wants to try to authenticate in the Client/Server architecture.

(ii)The service is ensuring to the client is accessing to authenticate with the server.

(iii)Kerberos Security Server had known as Key Distribution Centre (KDC) which is reliable for vice versa the user and the service.

(iv)The KDC works by storing Secret Keys (password) for the users and services and communicated with each other.

4.2 Apache Hadoop Encryption Security

Encryption can be used on different layers in a traditional data management software/hardware stack. If we choose encrypt at a given layer, it outcomes come with different advantages and disadvantages.

• Application-level encryption. It is fully protected and linear paradigm. The process has endmost control over what is encrypted. It reflects the requirement of the user. However, it is very difficult to write an application to do this. It does not provide an option for customers of existing application that do not support encryption is hard.

• Database-level encryption. It is same as an application level encryption in the terms of its properties. Generally, database vendors offer some forms of encryption. However, there are may be some issues for performance. It is an example that the indexes cannot be encrypted.

• File system-level encryption. This encryption has features of high performance, and deploy easy to place. However, it is not capable to structural some application-level mechanism. For instance, multi-tenant process try to encrypt which is based on the end user. But a database gives importance to different encryption settings for every column stored within a single file.

• Disk-level encryption. It is easy to deploy and high performance, but it is quite inflexible. In fact it prevents paradigm against the physical theft.

In this stack HDFS-level encryption is fit between database-level and file system-level encryption. It gives the positive effects. HDFS encryption is capable to maintain better response and previous Hadoop applications want to execute transparently on encrypted data. HDFS also has more context than traditional file systems when it comes to make policy decisions.

HDFS-level encryption also protects the attacks on the file system-level and below (so-called "OS-level attacks"). The operating system and disk only make interaction with encrypted bytes, because the data is already encrypted by HDFS[11].

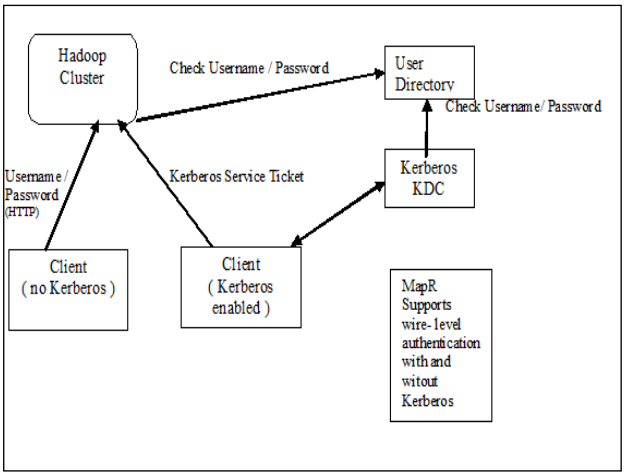

Figure 3: Hadoop Security Architecture

Step 1: Initial Data Encryption Level

- In the encryption method transferred and breaks the input 64 -bit into Right half and Left half.
- Acquires a left shift and is transferred (56 bit key)
- Substitution method is used to converted the right half key
- One more round of changing position takes Place Left half and the modified right half merge in the structure of new right half
- Previous right half converts in the new left half
- This loop continues in 15 more times

Step 2: The Hadoop Encryption

- Primarily, data key generated and it generates the key.
- Map Reduce job execute on the nodes in cloud and generates data.
- Data required to be encrypted with cluster configuration.
- Data key used for encrypting data as data output.
- Encrypted data and data key stored.
- Key- encrypting key and separately stored.

Step 3: Using Data Node symmetric Key for Decrypting the Data Block

- Client sends the request for encrypting data to name node. Trust stores with authenticate and key store with symmetric keys client -side authenticate and after that symmetric key is used for data access.

-Keystone with homogeneous keys and public key pairs /certificates.

-For comparing symmetric key, Name node authenticates the request is used for its own key.

-If the most importance a binding name node brings a list of nodes features and information.

-Key stores with symmetric keys and public keypairs/ certificates, client sends the request data blocks for encrypting data to data node.

-Data node uses the key to decrypt the data block and than successfully passes it back. Also, respective data nodes pass subsequent data blocks back.

-Data node communicates to decrypt an access subsequent of data block [5].

4.3 Data Encryption at Intel Hadoop Distribution with the use of HIVE

HDFS encryption works by absorbing all Hive information inside the same encryption area. Cloudera Manager is the Hive Scratch Directory to be private in encryption region.

- It creates symmetric key and keystores.
- It creates a key pair (private /public) and key store.
- It also creates Create a trust store contain public Certificates.
- By decoding certificates from a key store define in step 2 and import them into a trust store.
- Hadoop component "pig" is used the symmetric key to encrypt hdfs file.
- Hadoop component "HIVE" plays an encrypted external table uses the symmetric key that is created in step 1.
- Authorized client can access encrypted data

through map reducing jobs by using certificatesfrom trust store[8].

V. CONCLUSION

The aim of Big Data Anonymization is to protect the privacy of the end-user and make it legal to share the information while not obtaining permission from people. Big Data will provide the solution of new growth opportunities and the whole organization are related to those summations and analyze industry data. The goal and the phenomenon behind big data are growing rapidly, a comprehensive security component is needed to assuage risk of rupture and satisfy the best usage of big data technology.

REFERENCES

1. D.P. Acharjya, Kauser Ahmed P, "A survey on Big Data Analytics: Challenges, open Research Issues and Tools." (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 7, No. 2, 2016.
2. Karim Abouelmehdi, Abderrahim Beni-Hessane and Hayat Khaloufi, "Big healthcare data: preserving security and privacy" Published: 9 January 2018, Journal of Big Data2018 5:1.
3. Vanisha Mavi, Nidhi Tyagi, "Hadoop's Second Generation –YARN", International Journal of Contemporary Research in Engineering & Technology, Volume7, Issue 1, ISSN: 2250-0510, 2017.
4. Kanika , Alka and R.A. Khan, "An Improved Security Threat Model for Big Data Life Cycle." Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.7 No.1, 2018, pp. 33-39 © The Research Publication.
5. Ibrahim Abaker Targio Hashem,Ibrar Yaqoob , Nor Badrul Anuar , Salimah Mokhtar , Abdullah Gani, Ullah SameKhan "The rise of "big data" on cloud computing:", information system 22 July 2014.
6. W.Wang, K. Zhu, L. Ying, J. Tan, and L. Zhang, "Map task scheduling in mapreduce with data locality: Throughput and heavy-traffic optimality," in INFOCOM, 2013 Proceedings IEEE. IEEE, 2013, pp. 1609–1617.
7. Han J, Ishii M., Makino H. A hadoop performance model for multi -rack clusters. In: IEEE 5th international conference on computer science and informat ion technology (CSIT) 2013.
8. J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu, "Utility based Anonymization using local recoding. " Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 785-790.
9. Vanisha Mavi, Nidhi Tyagi, "Hadoop's Second Generation –YARN", International Journal of Contemporary Research in Engineering & Technology, Volume7, Issue 1, ISSN: 2250-0510, 2017.
10. Risha Tabassum, Dr. Nidhi Tyagi, "Hadoop Identity Authentication using Public Private Key Concept", International Journal of Engineering Trends and Technology, Volume-45, Number-9 -March 2017.
11. Vanisha Mavi, Nidhi Tyagi, " Data Compression Technique On Hadoop's Next Generation: Yarn", International Journal of Scientific Progress and Research, Issue 92, Volume 33, Number 01, 2017.