SECURITY SURVEY ON BIG DATA AND EFFICIENT ENCRYPTION METHODS

Vijayalaxmi Nandargi[1]

[1]PG Student, Acharya Institute of Technology, Bengaluru

Abstract

In the last few years, the growth in the data is enormous. Data is growing in the form of terabytes to Exabyte's, such data is difficult to manage and process it through traditional systems. One of the popular methods for processing large amount of data is Hadoop. Originally Hadoop system was developed without security models so security is major concern of Hadoop development. This paper concentrates on security issues in Hadoop and available solutions.

Keywords: *Big Data, Hadoop, Encryption, SASL, Kerberos*

---***---

1. INTRODUCTION

Big Data refers to the terabytes to Exabyte's. (Structured or unstructured) data. Such data is produced by social networking, medical records, industries, online transactions, and social media sites like face book, twitter, and also Smartphone's plays major role in growing volume of data. An application of relevance includes health care, retail consumer, finance and telecommunications, web and digital media, E-commerce and customer services. Big data plays a major role in research and analysis. Hence consists of millions of records from different areas. [1]

2. CHARACTERISTICS OF BIG DATA

- **Volume:** One of the aspects of Big Data is Volume. Due to increase in number of internet users the global data has been increasing exponentially. Nearly face book generates 500+ Terabytes of data per day, 42+ millions of like actions and more than 500 millions of photos per day. Conventional IT system faces volume as a challenge; Hence Organizations are overflowed with hundreds of terabytes of information.
- **Variety:** Variety defines to which category the big data belongs to, is also more essential fact to be known by the data analysts. With great growth in data sources there are different types of data which need analysis. It includes structured, unstructured and semi structured. Structured data like numbers, data, group of words, Unstructured like text, audio, videos and Semi structured like Non relational databases.
- **Velocity:** Velocity at this context refers to the speed with which data has been occurred, data has to be processed within a predefined time but traditional systems takes more amount of time to process. [2]

3. TRADITIONAL ENTERPRISE APPROACH

In traditional systems, Enterprises must have completely depended on database vendors like IBM, oracle to store huge amount of data because traditional system can't process such a large amount of data. Applications those practice less amount of data can be accommodate in a manageable database server. When it comes to process enormous amount of data. It becomes more difficult to handle and process such data through single database server.

To solve this problem, Google proposed a solution called Map reduce, using an algorithm called Map reduce. This algorithm divides the task into small parts and assigns them to computers, and collects the results from them which when integrated, form the result dataset.

Since, traditional approaches failed to process the large amount of Data. The Hadoop has two main components called HDFS and Map reduce. Architecture of Hadoop is shown in figure 1. These Software platforms allow writing and running applications easily, which makes easier to process large amount of data. [3]

4. ARCHITECTURE OF HADOOP

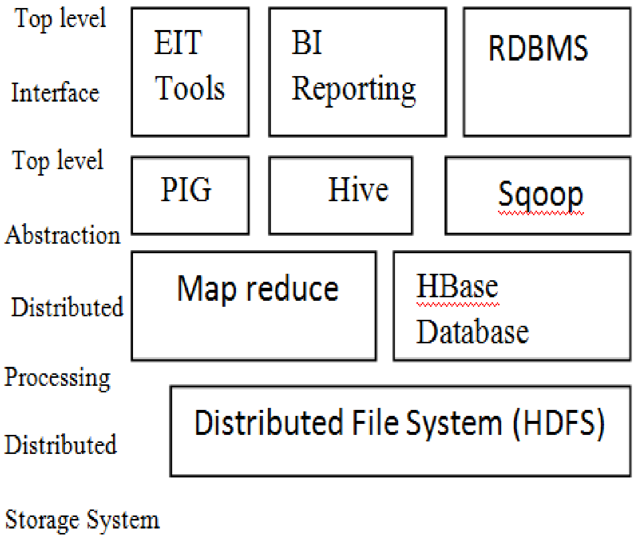

Fig 1: Architecture of Hadoop

- HDFS: It is responsible for storing large amount of data on the clusters. It is highly faults tolerant.
- Map Reduce: It is a parallel programming technique for processing huge amount of distributed data on clusters.
- HBase: It is NoSQL oriented. Used for random access of read and write.
- Pig: It is data programming language for analyzing data of Hadoop computations.
- Hive: A data warehousing application that provides a SQL like access and relational model.
- Sqoop: For importing and exporting relational database and Hadoop.
- Oozie: A workflow management and orchestration and for dependent Hadoop jobs.

4.1 Maps Reduce

Google has successfully used Map reduce. It's also used in Google for web services, mining, machine learning, etc. To analyse the hadoop data, a standard tools are shown in figure 2. The Map function performs a specific operation on each item. The Reduce function combines the items according to the designed algorithm. This model is easy to use even for the inexperienced people. It's highly fault tolerant, load balancing, provides maximum throughput.

Functions of Map reduce:

- Job client (client): Assigns a job to Map reduce
- JobTracker: That divides the subject task.
- TaskTrackers: A unit that physically performs the subtasks separated by JobTracker.
- Distributed file system: Used to share tasks files between multiple TaskTrackers.
- The input is provided to different map instances by run time device
- Map and reduce are user defined functions.

Standard tools to Manage Big Data:

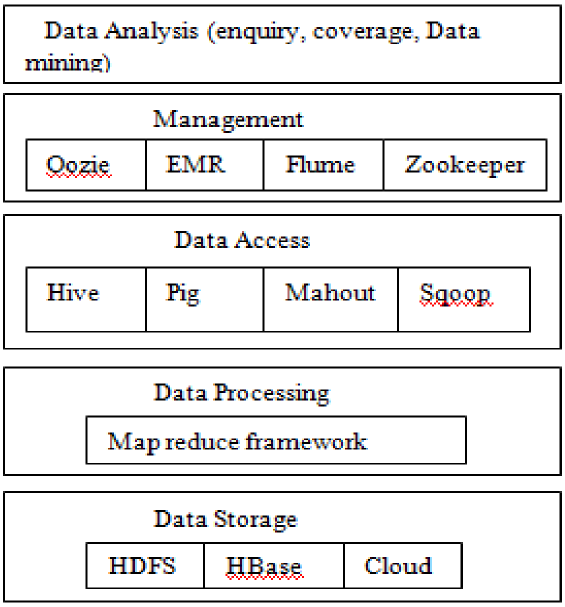

Fig 2: Big Data analysis tools

- Map reduce is used for processing Big Data. It is a framework that can handle large volume of structured and unstructured data in parallel.
- The tools like Hive, Pig, and Mahout are used to handle velocity and heterogeneity of Data.
- Oozie, Flume and zookeeper are used to handle veracity and volume of Big Data.
- These are the standard Big Data management tools. [6]

4.2 HDFS

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. As shown in figure 3. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant.

- Minimum charge for hardware and extremely fault tolerant.
- Provides high throughput access to application data, appropriate for applications of large data sets
- Enable streaming access to file system. [1]

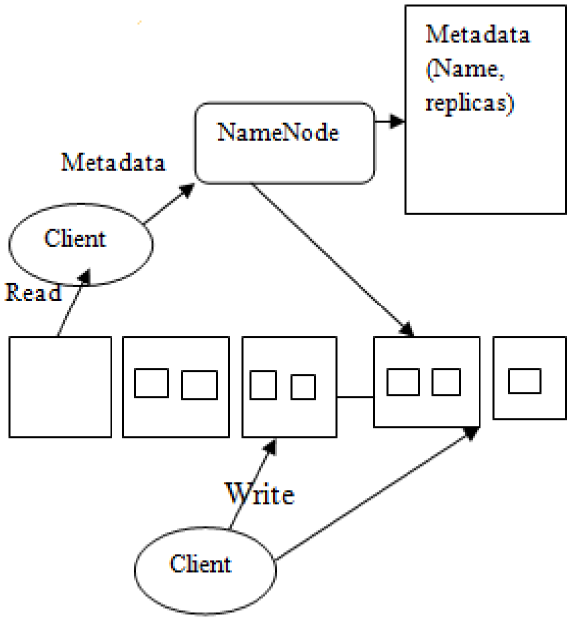

Fig 3: HDFS Architecture

HDFS includes NameNode and DataNode:
A **NameNode** is a one which is called as vital service node that manages the file system state. The locations of all the blocks are known by the NameNode. When a user wishes to retrieve a file from Hadoop, the information is retrieved via NameNode. This NameNode will indicate to the client about the blocks. The client then, recovers the desired blocks.

DataNode contains data blocks. They are initiated by NameNodes. DataNode responses to the clients based on their request. During the operations like reading and writing. It returns the location of the newly created blocks, while writing. Each DataNode is called upon the initialization of NameNode and also regularly, to return a list of stored blocks.

5. CHALLENGES IN BIG DATA

5.1 Data Isolation and Safety Measures

Security is one of the major challenge. Data has been growing exponentially which can't be always well understood by Companies and organizations. Also contains sensitive data of each individual. To reduce the privacy problem. The attributes which de-identifies the individual data need to be removed to secure each files. One need to follow this approach when data is released or managed and it only works correctly with the proper methods. Securing data is another major challenge in the time. These challenges consist of protection against security breaches and data leakage, penetrability in public databases, and third party data division. Big data provides a platform for users to store huge amount of sensitive as well as clumsy data. However secure data distribution is challenging. [5]

5.2 Heterogeneity and Incompleteness

Machine understands only homogeneous data. Hence, structured data is essential for efficient and accurate data analysis. Incorrect data analysis with incomplete data leads to load the whole process again and again. Consider a health record database that records, age, occupation and blood group for each patient. Many a time's, patients do not provide all the information and hence the value of such data is set to Null.

5.3 Relevance

The data is being available much faster in real time. The digitization makes new types of large and real time data over the years, So many organizations, industries have to process terabytes to pet bytes of data, usually it takes long time to analyse it. [3]

6. HADOOP SECURITY ISSUES

6.1 Securing Data in Movement

Since Hadoop stores data in distributed manner. So a message has to take place between Client and the DataNode. The user's data has to be transmitted across the network; by default this data is not encrypted. A variety of communication protocols like Transmission control Protocol over Internet Protocol, Hypertext Transfer Protocol (HTTP), and Remote Procedure Call (RPC) are used.

6.2 Securing Data at Rest

Data at rest refer to the data which is not accessed for a long time. Because of the nature of Hadoop. It stores data in the distributed manner. And Hadoop deals with huge amount of data and encrypting and decrypting process has to be fast. The Hadoop clients like IBM, Cloud era etc provides security to the users data. For this one should provide open framework where everyone can use it wisely and get secured Hadoop environment. [4]

6.3 Data Access Control

Unauthorised users can try to access the data; Third party users may hack the data, Proper authentication is need to save a user's sensitive data over the network.

7. EXISTING SOLUTIONS

Originally Hadoop was developed without much security. Security is a critical requirement. Initially authentication level was just with user name and password is provided by default because security was not the major concern. Default authentication is not provided by Hadoop. So to provide security to data some algorithms are designed. The Standard Algorithms are used for encrypting the data over the years.

7.1 Encryption

Converting the plaintext into unreadable from, saving sensitive data from unauthorized users.

There are numbers of methods used for Encryption
Symmetric encryption: The plaintext (file) has to be scrambled and convert it to unreadable form, while transmitting over the network, the file sender and receiver must use same key for both encryption and decryption.

Asymmetric encryption: It uses two different keys, Public key for encryption and private key for decryption. Best part is only one key need to be protected by the third party.

RSA Algorithm: It is an Asymmetric cryptography; RSA algorithm is designed by Rivest, Adi Shamir and Leonard Adleman, Hence the name **RSA**. It uses public key and private key for encryption and decryption. When clients and sever need to be communicated, the clients sends a file using public key and encrypt it. On the other hand server uses private key for the decrypting a file.

AES Algorithm: Advanced Encryption Standard is also a standard and most trusted algorithm by many organizations**.** It works efficiently using 128bits, It also uses key of 192 and 256bits, for the purpose of heavy data encryption.

Limitations of AES:

- Hadoop uses AES for encryption.
- AES requires more memory hence degrade the performance.
- AES is tricky and implementation requires more time.
- AES requires lookup time for every time.
- Since AES is complex and hard for implementing in software.

With all the above limitations Bern Stein proposed a stream cipher algorithm called Salsa20.Most reliable one is Salsa20. It supports both 128bits and 256bits.

Some security methods are available for securing the Hadoop data. Many algorithms like RSA, AES are used for encrypting data. In distributed file system, when two parties have to communicate with each other, initially they need to establish connection between them. The client approaches server to presents his password and it need to be verified by the server then makes the connection. Sometime password can reach unsafe network then there is threat for securing the data. Strong authentication has to be provided for clients and server applications by the Kerberos. The operation of kuber uses variant of Salsa20 called Chacha20 because of the number of reasons. Kerberos is an authentication protocol provides reliability, confidential to use data that sent between two parties. Chacha20 is designed such that, It can encrypt block of data, mainly suitable for block cipher. One more advantage is each individual blocks are decrypted by providing suitable key. For encryption process, the Hadoop clients configure the key in KMS. Then data is stored in HDFS, Initially data is encrypted at client side and divides into smaller chunks, at presently chacha20 is available for Kuber. Map reduce can be applied to each chunks and before processing decryption method is called.

7.1.1 Encryption of Data in Movement

Some underlying protocols are used when data is transmitted across the network in Hadoop. To secure the user data in action. A framework called SASL is more helpful. Client sends verification note to NameNode and JobTracker services is done during Hadoop's remote procedure call using the SASL framework. Kerberos is used as the authentication protocol to authenticate the users within SASL. To protect the data in motion, The Authentication framework called SASL is used to encrypt the data as it is being imported into the Hadoop ecosystem. The data exchanged between two parties is encrypted and security is given by SASL. It is not understandable by any hackers.

7.1.2 Encryption of Data at Break

Implementing encryption is not so easy, Hadoop has figured out number of choices for implementing encryption at break. Encryption at break indicates the data at rest, so to secure the stored data. Hadoop comes up with a distributed file system called HDFS. Encryption need to be very transparent so in encryption region the individual content is encrypted transparently upon write, and decrypted upon read. When a encryption region is created, each encryption region is coupled with a single encryption region key, and it is specified when the region is created. Each encryption region has its own single data encryption key (DEK). HDFS never handles these keys directly. As an alternative, HDFS only ever handles an encrypted data encryption key (EDEK). EDEK is decrypted by clients, to read and write it then use the subsequent DEK. HDFS is simply seeing a stream of encrypted bytes. [5]

8. CONCLUSION

Thus, in this paper we have concluded a security issues and efficient methods for securing data by encrypting, not allowing third party to catch the user sensitive data.

REFERENCES

[1]. Gole, Sheela, and Bharat Tidke. "A survey of big data in social media using data mining techniques." *Advanced Computing and Communication Systems, 2015 International Conference on*. IEEE, 2015.

[2]. Arora, Sanchita, et al. "Big heterogeneous data and its security: A survey." *Computing, Communication and Automation (ICCCA), 2016 International Conference on*. IEEE, 2016.

[3]. Menon, Sindhu P., and Nagaratna P. Hegde. "A survey of tools and applications in big data." *Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on*. IEEE, 2015

[4]. Jaseena, K. U., and Julie M. David. "Issues, challenges, and solutions: big data mining." *NeTCoM, CSIT, GRAPH-HOC, SPTM–2014* (2014): 131-140.

[5]. Parmar, Raj, et al. "Large Scale Encryption in Hadoop Environment: Challenges and Solutions." *IEEE Access* (2017).

[6]. Sharma, Priya P., and Chandrakant P. Navdeti. "Securing big data hadoop: a review of security issues, threats and solution." *Int. J. Comput. Sci. Inf. Technol* 5.2 (2014): 2126-2131.

[7]. Neuman, B. Clifford, and Theodore Ts'o. "Kerberos: An authentication service for computer networks." *IEEE Communications magazine* 32.9 (1994): 33-38.

[8]. Cuzzocrea, Alfredo. "Provenance research issues and challenges in the big data era." *Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual*. Vol. 3. IEEE, 2015.