EAI Endorsed Transactions
on Smart Cities

Review Article EAI.EU

Protecting and Securing Sensitive Data in a Big Data Using Encryption

Praveen S. Banasode[1,*], Sunita Padmannavar[2]

[1]Jain College of Engineering, Belagavi Affiliated to Visvesvaraya Technological University, Belagavi, India

[2]Gogte Institute of Technology, Belagavi, Affiliated to Visvesvaraya Technological University, Belagavi, India

Abstract

The Transaction data which contains a sensitive data, a program like a android app or a browser, does not adequately protect information such as unique values or related payment information, more or likely a privacy concern. In most of the cases, security breaches, which involve the unstructured data like documents and files, will reveal all sensitive information. To address this issue the transaction data can be processed across the nodes based on Advanced Encryption Standard(AES) algorithm for generating keys and also by using MapReduce algorithm to check number of sensitive data, where we will partition the data based on set key value pairs, whereby protecting the raw data using real-time security monitoring. The data, which requires an extra protection, needs to be identified, based on that data can be encrypted.

Keywords Advanced Encryption Standard (AES), real-time security monitoring, sensitive data.

Received on 05 March 2020, accepted on 08 April 2020, published on 17 April 2020

doi: 10.4108/eai.13-7-2018.163991

*Corresponding author. Email: praveenb.jce@gmail.com

1. Introduction

Big Data mining is the capability of extracting useful information from these large datasets or streams of data, to maintain the confidential and privacy concern of individual person and there data is not compromised.

Transaction data are often analyses to support, for example, personalized web search, however such data contains a sensitive information about individuals, releasing them in their original from may lead to privacy breaches[8].

Transaction services providers receive Exabyte of data from their customers, as these data has commercial and legal information the company needs to maintain for a specified period of time. Storing and managing the data is very challenging as it is growing minute by minute. The firm starts outsourcing the data for their profit this comprises the customers privacy.

In this study Map Reduce may be used to achieve, Protecting communications - data in transit should be adequately protected to ensure its confidentiality and integrity. And what are the problems faced by the clients and which improvements are required for the Big Data service provider. The above facts thus leads to the importance of this study in identifying the sensitive data thereby creating control policy, confidentiality and protecting the data[8].

2. Related Work

The distance based encryption technique is applied for in biometrics where the threshold value can be decrypted with private key of another key[2]. Many algorithms use encryption and decryption using a random key generator in IoT communications ASCII and XOR operations are used[4]. Data hiding of image, to hide extra information of different embedded layers the original image is encrypted with less loss of data[5]. Data is encrypted and decrypted within the database by using 256 bits of AES encryption, the data is stored in database confidentiality data is retrieved efficiently[11]. AES algorithm generates a new key for each input of image where it constantly changes key while encrypting the data[12]. Reverse engineering is applied for encrypted data analyzes plain text protocols uses pin to record executed

instructions[13]. Improvement of encryption data can also be implemented by using hybrid encryption algorithm[15]. An attribute based encryption(ABE) only the matching attributes can decrypt it where data is encrypted with AES and the AES key is encrypted with ABE[17].Two layer encryption is used for issues related to securing encryption algorithm for securing data[18].

The paper is organized as follows: Section 2 briefly outlines the backgroundDiscuses the main Objective. Section 3 provides a literature survey pertaining to security in Hadoop. Section 4 provides Section 5 Experimental results are presented. Section 6 concludes the paper.

3. Background

When a user enters any information on a web application, it is based on the trust that the server will protect that sensitive data, but the data breaches occurring time and again results in doubts the truth of the value, exposure of the sensitive user data has increased[16]. These are leading into a serious violations of data privacy and information security. Exposure of sensitive data can be caused by any form. An unauthorized person or third party poses more threat they have access to the information and have the power to sell it to the external users[7].

We have proposed a system based on a distributed parallel algorithm using MapReduce to identify pattern of sensitive data and non-sensitive data and its similarities. The MapReduce - based Search, Sort, intersection of algorithm are used to identify the sensitiveness of data. To process, analyze and visualize the data points, Building security in these applications from the beginning is beneficial in the long run[6].

4. Objective

The present research work was under taken with prime objective of identifying the various problems of Transactional data.

- Personal Information like SSN/SIN, DOB, AdharCardNumber.
- Banking Information like Account Number, Debit/Credit Card Number, ATM Pin Number, Registered Mobile Number.
- The main objective of this work is to protect sensitive data from threats using sophisticated secure algorithm in a distributed manner using Hadoop technology. Which takes care of the following parameters.
  - Efficiently storing and retrieving the bulk data.
  - Detect and Protect Sensitive information.
  - Reliable security mechanism
  - Robust and Fault tolerant system

An application encrypts Debit/Credit card number but also decrypts this data when retrieved. By designing a good algorithm we can protect the sensitive data so that the potential leaks can be controlled.

4.1 *Security Goal*

To prevent attacks, some measure are required for security implementation like:

- By Encrypting the data: It is very important to encrypt the data, where the data is in the form of plain text. By identifying the data which requires an extra protection and limiting the accessibility.
- Authentication: Encrypted data remains private by means of not disclosing the confidentiality of the user.

5. Method

A collection of data which is arranged in the form of sets, consisting of duplicate and unique values. There is more or likely a chances of duplicate data in sensitive data. By using the set intersection we can eliminate the duplicate value from sensitive data set.

Consider a set which is consisting of user information like Name, SSN, Debit_Credit_Number, Email and in another set of sensitive a data of SSN, Debit_Credit_Number sequence of data, like that many information is carrying,

let us consider, two sets s1 and s2.

s1={{abc,123,1234567890,abc@m.com},{bcd,234,2345612312,bc@g.com},{azx,345,0987654321,zx@g.com},{wer,347,2314560987,x@g.com},…..}

s2={{bcd,234,2345612312,bc@g.com},{wer,347,2314560987,x@g.com},{abd,389,4534567890,abc@m.com}, ... }

Each object is called an element of set. For two sets s1 and s2 is the number of elements present either of the set (s1Us2). total number of elements present in both the sets s1 and s2 (s1∩s2)[1].

The sets which contain all the elements of a given collection is called the universal set which is represented by 'μ' [1].

$$\mu(s1 \cup s2) = \mu(s1) + (\mu(s2) - \mu(s1 \cap s2))$$

Definition: Let s1 and s2 be arbitrary given sets. By function f:s1->s2 from the set s1 into s2, a rule which assigns to each member x of X, a unique member f(x) of s2.

The member f(x) is called data of x under the function(mapping)f or the value f at x. The set s1 is called the domain of f and the Set s2 is called the co domain of f.

The set of element f(x), x ε s1 is called range of f. Thus, the range of f is a subset of s2.

Let f:s1->s2 be a mapping from the set s1 into the set s2.

If f(x1)=f(x2)=>x=x2 for every x1,x2εs1, if and only if the data of distinct points s1 are distinct, that is x1≠x2=>f(x1)≠f(x2).

Thus, from this definition we arrive to conclusion the data is protected and thereby removing the duplicates.

Same methodology of set theory will apply in processing chunks of data into the HDFS cluster the above sets are examples of unique identifier as well how these identifiers are carrying the sensitive data for processing the data from one end to another end and in between many transformation will take place and also how the real time data is protected and secured using encryption and decryption process.

Programs can directly access key value pair matching index and even made available on the network. The MapReduce algorithm is implemented to restrict information getting leaked to unknown sources using two interfaces as Mapper and Reducer by manipulation so the result will not violate the privacy of the clients. Sample code is presented for preserving sensitive information of client.

Algorithm:

```
class Sensitivity:
def
Mapper(key,_,value):(name,Ssn,debitnumber,email)=line.
split(‘,’)
yield debitnumber, int(Ssn)

def Reducer(key, debitnumber,Ssn): totalnumber=0
numberdebit=0

for x in Ssn:
   totalnumber+=x
   numberdebit+=1
   yield debitnumber
```

The data which is used to store more memory requires to be sliced to store in proper format and also it helps the functions to share widely on distributed platforms for any transactional purpose. The sensitive data is used wisely by checking all possible threats from the outside world with proper algorithm to process the data in a way it required for successful transactions.

Each tuple of value is having a multiple value for DebitNumber and CreditNumber, so this increases the task of protecting the data of a user in its best interest and also to be processed efficiently. By decomposing the data into two relation we can further reduce the cluster of work which stored on disk. Some users do not have credit card number for them NULL values is assigned, they may not have or may not be available for the time being in that case a NULL is stored for such attributes.

Multivalued are decomposed into atomic value by properly arranging primary key and other keys. Decomposing the data helps in experimenting the data in safer and we can have multiple set of data which is going to be more efficient. A mechanism is employed to secure the data which combines encryption and decryption when such things are implemented in the algorithm then framework need not do changes.

Table 1. a) Nested Relation of attributes within each tuple.
b) Normalised attributes of relation into two different relation.

a)

Ssn	Name	DebitNumber	CreditNumber
123	abc	1234567890 7894561231	4563217890
234	bcd	2345612312 7456321456 1236547890 6611223355	5003412345 5500223344
456	qew	3344552200 8880001234	NULL
639	asd	7452316970 5544332211 5007800901 9600120013	4563214600 7452130011 9004455333

b)

Ssn	Name

Ssn	DebitNumber	CreditNumber

The collected information will be converted into a secrete key code which will not be understood until it is decrypted by hiding true information of its meaning. An advanced encryption standard (AES) is the symmetric key which uses of government standards of classified information. The key shared among system and the computing environment which uses the key to decrypt.

```
import AES import base64 import os
def data_encryption(privateInfo): BLOCK_SIZE = 32
padding = '{'

pad = lambda s: s + (BLOCK_SIZE - len(s) %
        BLOCK_SIZE) * padding
Encodedata=lambda c, s:
        base64.b64encode(c.encrypt(pad(s)))
```

```
secretkey = os.urandom(BLOCK_SIZE)
print 'encryption key:',secret

cipher = AES.new(secretkey)
encoded = Encodedata(cipher, privateInfo)
print 'Encrypted string:', encoded
```

The unauthorized person is not be able to access any data which is in an encrypted form, the sensitive data is made sure secured by encryption algorithm, all the digital data is stored and transmitted over the computer network.

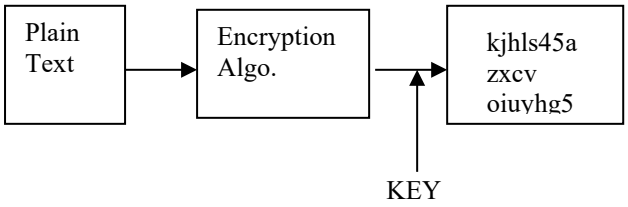

Figure 1. Encryption operation

Architecture of computation[14]

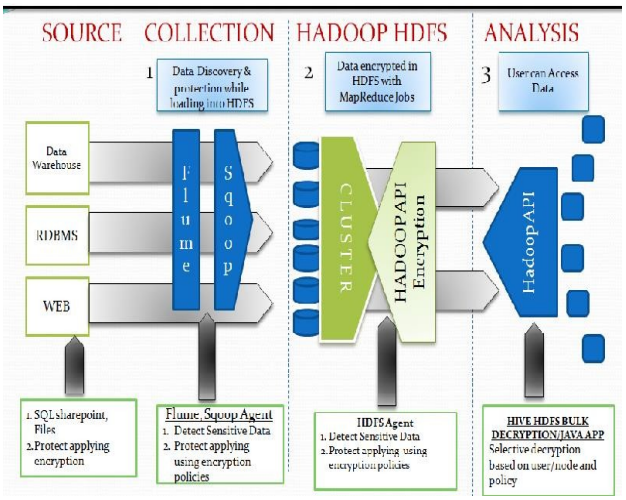

Figure 2. Architecture Computations

i. Focus is on real time data is, for new level of complexity to derive value from data. The margin of detecting and dealing set of data is vanishing and also risks around big data over time it will shift by focusing on the tooling system for storing and addressing in a more secured way by implementing the MapReduce algorithms. In short risks will perish using the above methodology.

ii. Failure of data is very less, here by applying the encryption policies for the data which is passed through some channels protects the sensitive data collected from the source.

iii. Duplication of data is reduced because the same data is collected in clusters on unique values which are been encrypted and during moving it may be that the same source data is required to process on multiple data stores and consuming applications, and these operations will give rise to expected results.

iv. By using the selective descriptive analysis the real time data which is protected and secured in cluster will be decrypted based on use node policy and thereby the end user can access the data for further processing.

v. Similarly, the real time data which is available in abundance is a big challenge to control and secure the data and eliminating the duplicates by performing a proper set of analysis.

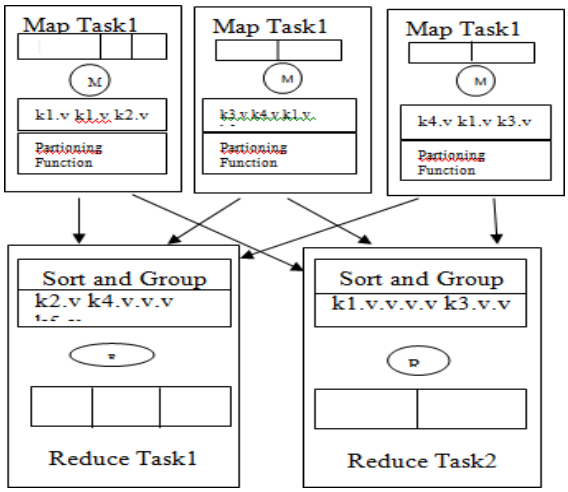

Figure 3. MapReduce Distributed Computing

MapReduce and Hadoop can partition this job and breaking up of this data into mappers most of them running parallel on different computers so imagine you have entire clusters of mappers and each responsible parsing chunks of data for input data and outputing the key value pair <k,v> for short cutting the data and then similarly that can be sorted and grouped together and reducer also run on one machine.

Now putting the result into result front machine and combining all together at the end. This illustrates how can the divide and conquer works on large dataset using MapReduce and Hadoop to process datasets on an distributed computers. To process data which even unfit in personal computer machines on distributed task over cluster of wide computer system to achievable task by taking up large datasets. That is how the MapReduce

works for scaling up to the BigData on distributed environment.

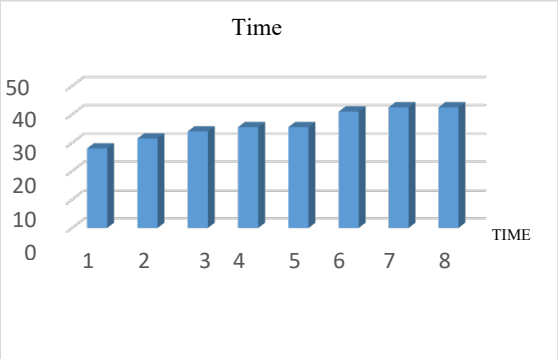

Figure 4. Time Efficiency on large data Set

The experiment demonstrates that difference in timing is negligible, if more data are tested the time efficiency decreases exponentially with increase in testing data of reduced value to its sensitivities.

6. Conclusion

The bad data can be detected and dealt, and also margin of errors can be reduced thereby shaping up of data. The data quality can be addressed to the point of data acquisition by ensuring the data in terms of consistency and accuracy. The data is made sure end-to-end encrypted and secured. The algorithm will collect data-quality measures like dealing up of data drift and sensitive data for transaction processing on a distributed cluster of computers the key value pair is made sure the data is encrypted and data will be decrypted only to the destination. The mechanism helps the framework to be more robust and need not require much do changes and also secures the safety of data by giving more transparency and meaning to the data. Name and value attribute have addressed the security concerns of data. The work is limited to a certain disjoint attribute and can be enhanced to large scale deployment by use of hybrid and public cloud services.

References

[1] Babu Ram “Engineering Mathematics” set theory and functions, pearson publications.

[2] Fuchun,Willy,Yi "Distance-Based Encryption: How to Embed Fuzziness in Biometric-Based Encryption" IEEE Transactions on Information Forensics and Security (Volume: 11 , Issue: 2 , Feb. 2016).

[3] Jindan Zhang, Xu An Wang, Jianfeng Ma "Data Owner Based Attribute Based Encryption" 2015 International Conference on Intelligent Networking and Collaborative Systems.

[4] Iqra Hussain, Mukesh, Nitin "Proposing an Encryption/ Decryption Scheme for IoT Communications using Binary-bit Sequence and Multistage Encryption" 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO).

[5] M. Hemalatha, K. Rajasekhar "Multiplicative symmetric key generation based encryption for data hiding" 2017 2nd International Conference on Communication and Electronics Systems (ICCES)".

[6] Meiko Jensen "Challenges of Privacy Protection in Big Data Analytics" 2013 IEEE International Congress on Big Data.

[7] Prof. Elisa Bertino "Big Data Security and Privacy " 2016 IEEE International Conference on Big Data (Big Data).

[8] Quang Tran and Hiroyuki Sato "A Solution For Privacy Protection In MapReduce" 2012 IEEE 36th International Conference on Computer Software and Applications.

[9] R.Manjusha and R.Ramachandran "Comparative Study of Attribute Based Encryption Techniques in Cloud Computing "International Conference on Embedded Systems - (ICES 2014).

[10] Roger Schell " Security – A Big Question for Big Data" 2013 IEEE International Conference on Big Data.

[11] Than Myo Zaw,Min Thant "Database Security with AES Encryption, Elliptic Curve Encryption and Signature" 2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF).

[12] Vasyl, Ivan, Roman, Victoria "Information Encryption Based on the Synthesis of a Neural Network and AES Algorithm" 2019 3rd International Conference on Advanced Information and Communications Technologies (AICT).

[13] Wei Lin, Jinlong Fei, Yuefei Zhu, Xiaolong Shi"A Method of Multiple Encryption and Sectional Encryption Protocol Reverse Engineering" 2014 Tenth International Conference on Computational Intelligence and Security.

[14] Web Resources “Fig. 5.2 Architecture Computation”, Encryption achieved through filters.

[15] Xinqiang, Lili Yu,Lihuan "The application of hybrid encryption algorithm in software security"2013 3rd International Conference on Consumer Electronics, Communications and Networks".

[16] Xinhua Dong, Ruixuan Li, Heng He, Wanwan Zhou, Zhengyuan Xue, and Hao Wu " Secure Sensitive Data Sharing on a Big Data Platform" TSINGHUA SCIENCE AND TECHNOLOGY IS SNll1 007 - 0214ll0 8/ 1 1llp p 72- 80 Volume 20, Number 1, February 2015.

[17] Yanli Ren, Shuozhong Wang, Xinpeng Zhang, Zhenxing Qian "Fully Secure Ciphertext-Policy Attribute-Based Encryption with Constant Size Ciphertext " 2011 Third International Conference on Multimedia Information Networking and Security.

[18] Yoshiko,Hiroki,Hayato "Attribute-based proxy re-encryption method for revocation in cloud storage: Reduction of communication cost at re-encryption" 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)

[19] Zainab, Mahmood "New Fully Homomorphic Encryption Scheme Based on Multistage Partial Homomorphic Encryption Applied in Cloud Computing"2018 1st Annual

International Conference on Information and Sciences (AiCIS).