IJCSC 0973-7391 Volume 7 • Number 2 March 2016 - Sept 2016 pp. 90-94 available online at www.csjournalss.com

Big Data Security and Privacy Using Data Transformation with Role Based access Control

Shivani Patel, Het Prajapati, Prof. Ketan J. Sarvakar
Mtech (Information Technology), U.V. Patel College of Engineering, Gujarat

Abstract-Big Data refer to the analysis of significantly large collection of data that may contain user data, sensor data or machine data. It consists of data sets that are of large magnitude (Volume), large collection of data which diverse representation include structured, semi structured, or unstructured data (Variety), and should arrive fast (velocity). Big data implies performing computation and database operations for massive amounts of data. Security and privacy of Big Data is a big concern. There is no business or industry which is not involved in solving Big Data security and privacy problems. A variety of Big Data security and privacy techniques compete for the best solution. It is critical task to provide security and privacy for Big Data. This Paper aims to provide security and privacy to Big Data using Role-based access control technique. In Role-Based Access Control (RBAC) technique, access decisions are based on the roles that individual users have as a part of an organization. The other technique is introduced in this paper that is Data Transformation Technique. In Big Data there can be multiple files with effective names which all user can understand so in data transformation technique file names will be replaced by some code words or applying hashing technique which only particular user can understand which is going to use that data. Moreover, swapping and hashing of data is used in every file which are generated. Authentication is also used to provide confidentiality.

Keywords- *P*rivacy preserving data security; data transformation; role based access control; swapping; hashing.

I. INTRODUCTION

Big Data is a collection of large data sets that are complex in nature. There are both structured and unstructured data that grow large so fast so that it cannot be manageable by traditional relational database systems (E.g., RDBMS) [1]. As we know on daily basis data is increasing so fast so it needs strong business intelligence to provide privacy and security to data. It is hard to decide which data we should store and which data we should discard due to the large size of the data. So decision making becomes tougher due to the data required size. It is also big question that how to store large data that can be accessible easily.

If we see the facts available, in 2011 about more than one trillion GB of data was produced, and by 2020 it is expected that data will grow 50 times more than 2011. According to analysis Google receives more than twenty lac searching queries in a single minute and on YouTube seventy two hours of video are added every minute. Moreover, in every minute 217 new mobile internet users are added, and if we see for the social media then in every minute twitter users send over 1lac tweets and in every single minute brands are getting more than thirty thousand likes [1]. Security and privacy for this kind of data is big concern.

Related Work: The main objective of privacy preservation is ensuring that private data remains protected, while processing or releasing sensitive information. Privacy concerns about data have been raised in various literatures [2], [3], [4]. S. Moncrieff et. al. [5] proposes a solution to dynamically alter privacy levels based on environmental context using data masking techniques to decrease the intrusive nature of the technology, while maintaining the functionality. S. Meyer, ET. al. [6] demonstrates selected information discloser through a privacy manager module for a context-aware system interacting with a user. S. Bagüés et. al. [7] proposes a framework to control the dissemination of data within the context-aware service interaction chain, based on a set of user defined privacy policies. G. Drosatos et. al. [8] introduces a privacy preserving cryptography approach for distributed statistical analysis of data. All of the discussed solutions however, are very specific and address privacy concerns required for their solutions.

Our Contribution: This paper presents an approach for big data security and privacy. Role based access control is used to provide accessibility and privacy over data. Hashing and swapping of data is used to hide the data from adversaries. Authentication is also used to provide confidentiality.

II. PROPOSED SOLUTION

In this paperwe have proposed solution to the problem of Big Data security and privacy. In which there are different techniques which will be used for protecting Big Data while preserving privacy also.

Data transformation Technique: The aim of this transformation is to achieve isolation from sensitive information on the data used for any processing. It enables configuration of all attributes that are deemed sensitive and have potential to reveal privacy. The identified data attribute will be hashed and swapped. The non-identifiers such as timestamp and value elements remain unhindered. Data transformation algorithm. Here for example we are applying this technique on geography data and in that Backforce column will be transformed.

Algorithm: Data Transformation
Input : int i,j,tc, float temp
Output : Transformed data

```
1 Set tc = total number of rows;
2 Set i = 1;
3 If (tc mod 2) !=0 then
4 Set tc = tc-1;
5 End if
6 While (i*2) < tc do
7 Set j = tc – i;
8 Set temp = Backforce(ith row);
9 Set Backforce(ith row) = Backfroce(jth row);
10 Set Backforce(jth row) = Backforce(temp);
11 Set i = i+1;
12 End while;
13 if i = tc then
14 Set temp = Backforce(ith row);
15 Set Backforce(ith row) =  Backforce((tc)th row);
16 Set Backforce((tc)th row) = Backforce(temp);
17 End if
18 set i = 1;
19 while (i<=tc) then
20 Hash (Backforce);
21 set I = i+1;
22 End While
23 End
```

Moreover, Role Based Access Control will also apply for access permission over data. This module aims to provide access to the system through suitable mechanisms that fulfil access control requirements. When user will authenticate then after authorization it provides the user different privacy level according to set of rules. After a user is authenticated, the module then provide a list of hashed and actual primary identifiers which data the user requested and is authorized based on a set of rules generates. It also determines the level of privacy preservation for shared results must fulfil [9],[10]. User having same authorization may have different privacy level also for an example personal doctor may have complete authorization for their patients without hiding personally identifiable information and specialist doctor may have all the authorization but personally identifiable information may be hidden. The same is true for other users such as nurses, researchers. So sometimes higher authority may have lower privacy level also it only depends on the user is how connected with the data. So it's not only depends on the authorization but it also depends on the privacy level of particular user.

III. IMPLEMENTATION WORK

Here, we have shown swapping and hashing of data. We have taken geography data as sample data. We have swapped Backforce value first after that we have applied hashing of that data. Here we are swapping data for providing privacy and hiding the data from adversaries. Hashing is applied for security concern. We are using MD5 algorithm for hashing purpose. In our data there are three roles: Mining engineer, geologist and general user. So for

role based access control we have given all access permission to geologist and mining engineer. For general user we have not given access control over data so they will get only transformed data. In future we will apply more rules for role based access control.

FileName	Depth	Backforce
MINE_X_POINT _70_100.csv	90	3.49507
MINE_X_POINT_70_100.csv	100	3.6397204
MINE_X_POINT_70_20.csv	0	0.537588
MINE_X_POINT_70_20.csv	10	1.3336629
MINE_X_POINT_70_20.csv	20	3.8853729
MINE_X_POINT_70_20.csv	30	1.2258095
MINE_X_POINT_70_20.csv	40	1.1626911

Swapping Process

FileName	Depth	Backforce
MINE_X_POINT _70_100.csv	90	1.0900575
MINE_X_POINT_70_100.csv	100	4.6136665
MINE_X_POINT_70_20.csv	0	1.2042842
MINE_X_POINT_70_20.csv	10	2.350018
MINE_X_POINT_70_20.csv	20	3.3667662
MINE_X_POINT_70_20.csv	30	4.244505
MINE_X_POINT_70_20.csv	40	0.4941535

Hashing Process

FileName	Depth	Backforce
MINE_X_POINT _70_100.csv	90	0608d2f96c42eac5e4a4a633eeb232c1
MINE_X_POINT_70_100.csv	100	1ebc6a293407218e655262e9b86d564a
MINE_X_POINT_70_20.csv	0	62c112a2c6303a85cd975157c67a197f
MINE_X_POINT_70_20.csv	10	b6b7ceb5ce6b3dbcc66b868d785cc218
MINE_X_POINT_70_20.csv	20	51367a43f5cb11c5b6d1de42e608cf39
MINE_X_POINT_70_20.csv	30	5deb993acc02cfc784ec570b5767a7c5
MINE_X_POINT_70_20.csv	40	6643beeda063d81ceb9e5a9fe49b9ba1

Fig. 1. Geography Data Example Dataset

We are using Hadoop tool for Big Data Storage and Processing purpose.There are number of software tools for processing and analysis of big data framework. Now a day's hadoop is most widely used tool for big data storage and processing. MapReduce technology using by google have been implemented by hadoop. It provides data paralleling automatically and also provide processing of data in computer. Apache developed many hadoop components and because of that are open source we can get them easily.

HDFS (Hadoop Distributed File System): it is the root of the Hadoop. It provides storage to the large data in the warehouse of hadoop. It also manage the storage in distribute format. It divides whole data which is given as an input into blocks and store it in the server in distributed manner. The concept used for conversation purpose is TCP/IP. It is fault tolerant so that breakdown of any element does not affect the whole system working.
Hive: Hive is mainly a data repository. With the help of hive row data can be stored in structure form via hdfs.
Components like **Sqoop and Flume** are used to transfer data between hadoop distributed file system and relation database like MySQL. Fig.2.shows the workflow of Hadoop processing which we have followed during implementation.

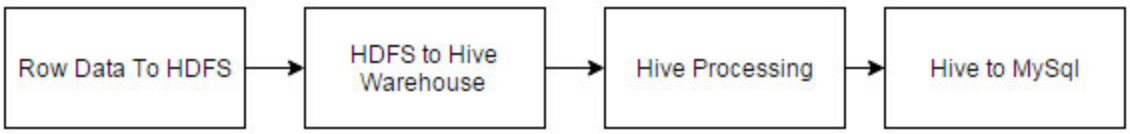

Fig. 2. Hadoop Processing Workflow

IV. EXPERIMENTAL EVALUAION

Here we are using correlation coefficient for evaluation of the technique we have applied. Correlation coefficient is used to find the relationship between data whether it is strong relation or not. Correlation coefficient value should be between 1 to -1. If the value of correlation coefficient is 1 that means when one variable's value is increasing by 1 then the other variable's value is also increasing positively. If the value of correlation coefficient is -1 that means when one variable's value is increasing by 1 then other variable's value is decreasing negatively. If the value of correlation coefficient is 0 then there is no relation between variables. When correlation coefficient value is nearest to +1 or -1 it shows strong relationship between data. Formula for correlation coefficient is given below. Here x represents original dataset and y represents swapped dataset.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} \quad \text{(Formula No. 4.1)}$$

Correlation Coefficient for the data we have taken is -0.68232048. It is nearly -1 and it is showing that it is good relationship between data. So it indicates better privacy with this method.

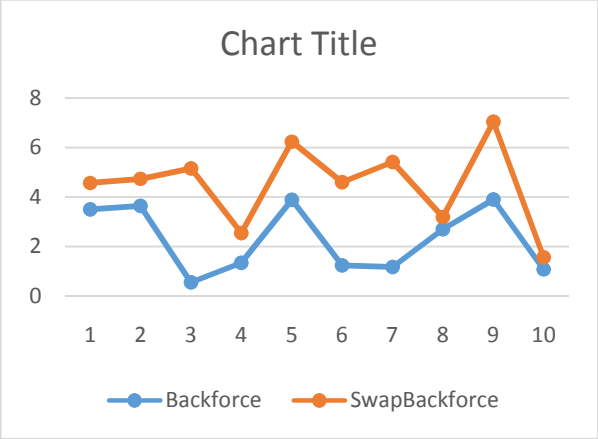

Fig. 3. Comparison between original and swapped data

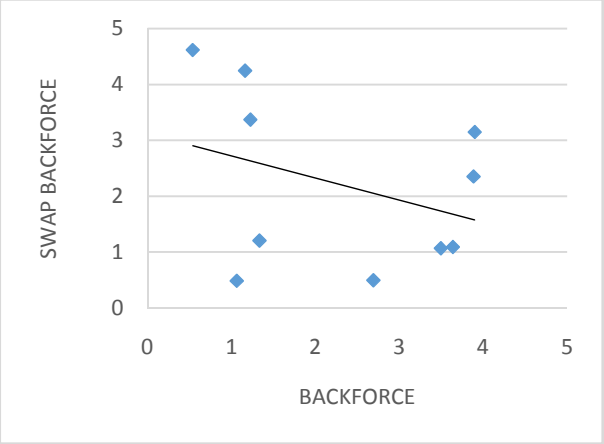

Fig. 4. Correlation coefficient indication graph

Fig.3. indicates relationship between original and swapped backforce. Fig. 4. Indicates that the line is decreasing so correlation coefficient is going negative.

V. CONCLUSION

We proposed an algorithm for big data security and privacy using data transformation technique. So far, the technique proposed focuses on modifying the database which involves scanning of database. The proposed algorithm carries out pre-processing of dataset using Hadoop and swapping and hashing of the sensitive attributes and identifiers. We worked on modifying the data such a way I can be hidden from adversaries. We have applied role based access control for giving access permission over dataset. We have obtained feasible results on dataset in terms of correlation coefficient.

VI. FUTURE WORK

The approach is easy to implement. Future work can include trying other hashing techniques and other ways of swapping the data. One can focus on a different rules for access permission.

REFERENCES

[1] "Big Data: The next frontier for innovation, competition and productivity"
[2] M. Chan, E. Esteve, et. al., "A review of smart homes—Present state and future challenges," *Computer Methods and Prigramns in Biomedicine*, vol.91. no.1, pp.55-81, Jul. 2008
[3] K. Courtney, G. Demiris, et. al., "Needing smart home technologies: the perspectives of older adults in continuing care retirement communities," *Informatics in Primary Care*, vol.16, pp.195-201, 2008
[4] G. Demiris, B. Hensel BK, et. al, "Senior residents' perceived need of and preferences for smart home sensor technologies," *Int J TechnolAssess Health Care*, vol.24. no.1, pp.120-1024, 2008
[5] S. Moncrieff; S. Venkatesh, et. al., "Dynamic Privacy in a Smart House Environment," *IEEE International Conference on Multimedia and Expo*, pp.2034-2037, Jul. 2007
[6] S. Meyer, A. Rakotonirainy, "A survey of research on context-aware homes," *Australian Computer Society,* vol.21, pp.159-168, 2003
[7] S. Bagüés , A. Zeidler, et. al., "Sentry@Home - Leveraging the Smart Home for Privacy in Pervasive Computing," *International Journal ofSmart Home,* vol.1, no.2, Jul. 2007
[8] G. Drosatos, P. Efraimidis, "Privacy-preserving statistical analysis on ubiquitous health data," *8th International Conference on Trust, Privacyand Security in Digital Business*, *Springer-Verlag,* pp.24-36, 2011
[9] R. Sandhu, E. Coyne, et. al., "Role-Based Access Control Models," *IEEE Computer,* vol.29, no.2, pp.38-47, Feb. 1996
[10] G. J. Ahn, "The RCL 2000 language for specifying role-based authorization constrains," *Ph.D. dissertation, George Mason University,Verfinia,* 1999