Access Control Based Privacy Protection Mechanism In Knowledge Extraction

[1]Varsha Gaur, [2]Mala Dutta
[1]Lecturer, [2]Assistant Professor
[1]Department of Information Technology, Prashanti Institute of Technology & Science, Ujjain
[2]Department of Computer Engineering, Institute of Engineering & Technology, Indore

***Abstract*— Privacy is a state in which one is not observed or disturbed by other people" Privacy protection policy is an approach to isolate the sensitive information from unauthorized access. Cloud security is one of the widely known fields in research. Cloud security focuses on having variety of security features like confidentiality, authentication, non-repudiation, access-control and integrity. In proposed framework the food mart dataset is the sample dataset which consist of thirty seven different tables to manage fourteen different departments. It demonstrates the basic working style of super market and schema about the storage of information. In our work privacy is focused with the help of algorithm named as k-means clustering with encryption. Along with it, access control policies are also applied named as role based access-control and attribute based access-control. The aim of our work is to reduce the searching computation time by minimizing the desired string and to obtain accurate results. The complete framework has been implemented in Java technology and tested on the basis of execution time and recall parameter.**

Keywords—RBAC, ABAC, K-means clustering, Encryption, Sensitive data

I. INTRODUCTION

The enhancement in technology is changing the practice of human. Development of industry without computer and use of computer without internet is a joke today. Internet based services and applications are rapidly emerging and increases demand to upgrade applications and existing solutions. Internet based large storage and services is known as cloud computing. Cloud computing services often rely on specific systems such as Hadoop Map Reduce, an open source proposed by Google. Map Reduce is being adopted by many academic researchers for data processing in different research areas, such as high-end computing, data intensive scientific analysis, large scale semantic annotation and machine learning.

As the rate at which we generate data increases, we find a greater and greater need to handle voluminous amounts of data within traditional machine learning algorithms. As a general rule of thumb, the more examples you can provide to a machine learning algorithm, the better it will be able to perform. The ability to quickly and efficiently process large amounts of data is necessary in order to effectively scale learning algorithms to match the growth of data available. Here we explore algorithms to keep privacy during knowledge extraction from large dataset.

With the increase in the volume of data, the demand for cluster computing has grown as problem sets become 50 larger and more interest develops in the field.

Along with this, time and speed has become the major factor in computing such data. This huge volume of data available needs to be organized and managed so as to facilitate proper understanding. Therefore the need of clustering comes into existence. Clustering analysis has been an emerging research issue in data mining due to its variety of applications. Its expensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has lead to the popularity of this algorithm.

A big challenge of preserving privacy and security in cloud computing is that developers and users wanted to spend as little effort and system resources on security as possible. Therefore, motivation of this research is how to design a system that satisfies below demands.

An environment context using data masking technique solution is proposed by S. Moncrieff [3] to dynamically alter the privacy levels in smart homes.

A solution proposed by S. Meyer [6] based on context aware system interacting with user is also proposed.

A solution proposed by G. Drosatos [7] for distributed statistical analysis of data from wearable source based on cryptography approach.

II. PROBLEM STATEMENT

"Privacy is a state in which one is not observed or disturbed by other people" Privacy protection policy is an approach to isolate the sensitive information from unauthorized access. To balance the competing goals of a permissive programming model and the need to prevent information leaks, the untrusted code should be confined. Traditional approaches to data privacy are based on syntactic anonymization, i.e., removal of "personally identifiable information" such as names, addresses, and Social Security numbers. Unfortunately, anonymization does not provide meaningful privacy guarantees and easy to reverse in many cases. These events motivate a new approach to protecting data privacy. In this paper we use access control policies with K means clustering algorithm with encryption for privacy protection. K-means is very quick, versatile, robust and easier to implement and deploy.

III. PROPOSED METHODOLOGY

Step 1: Data Collection

User and product is stored in dataset.

Step 2: Access Control Module

This module authenticates, authorizes and determines the level of privacy for any data share. RBAC and ABAC access control policies are used.

Step 3: Result Processor Module

This module control end users access to data processing results and ensures that privacy of any shared results is preserved.

The methodology also sketched in Fig. 1.

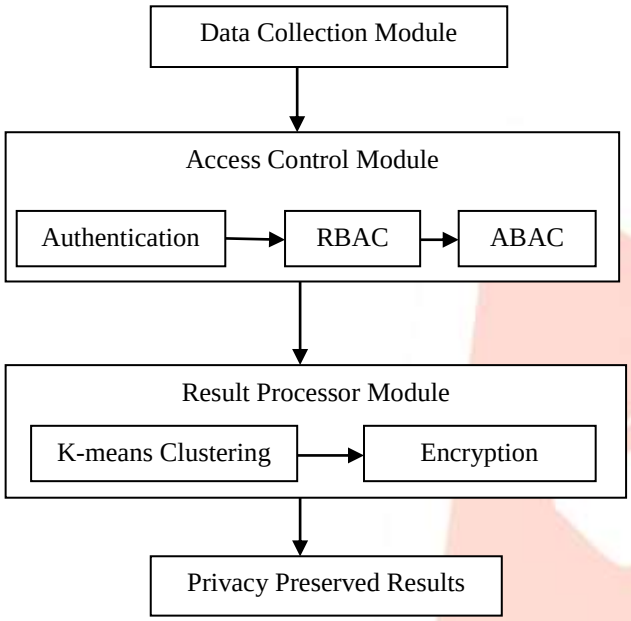


Fig. 1: Proposed System Architecture

IV. Data Collection Module

Data collection can be done in two ways either from a direct user or from the existing system. The auxiliary information can be acquired from a prior framework or dataset. We consider the standard food mart dataset as the prime source for evaluating purpose. This dataset can be downloaded from the Google using the following link:

http://pentaho.dlpage.phi-integration.com/mondrian/mysql-foodmart-database

It has been imported into MySQL database server. The food mart dataset is the sample dataset which consist of thirty seven different tables to manage fourteen different departments. It demonstrates the basic working style of super market and schema about the storage of information. A brief detail of the dataset has shown in Fig. 2

Table	Size
account	3.5 KiB
agg_c_10_sales_fact_1997	1.4 KiB
agg_c_14_sales_fact_1997	7.1 MiB
agg_c_special_sales_fact_1997	7.1 MiB
agg_g_ms_pcat_sales_fact_1997	277.3 KiB
agg_lc_06_sales_fact_1997	139.6 KiB
agg_lc_100_sales_fact_1997	4.7 MiB
agg_ll_01_sales_fact_1997	5.2 MiB
agg_l_03_sales_fact_1997	668.1 KiB
agg_l_04_sales_fact_1997	3.8 KiB
agg_l_05_sales_fact_1997	6.3 MiB
agg_pl_01_sales_fact_1997	6.2 MiB
category	3.1 KiB
currency	4.0 KiB
customer	2.6 MiB
days	1.1 KiB
department	2.4 KiB
employee	200.1 KiB
employee_closure	266.1 KiB
expense_fact	169.4 KiB
inventory_fact_1997	360.9 KiB
inventory_fact_1998	630.8 KiB
position	3.2 KiB
product	276.0 KiB
product_class	6.7 KiB
promotion	164.8 KiB
region	6.9 KiB
reserve_employee	23.5 KiB
salary	1.5 MiB
sales_fact_1997	7.2 MiB
sales_fact_1998	13.7 MiB
sales_fact_dec_1998	1.5 MiB
store	6.7 KiB
store_ragged	5.7 KiB
time_by_day	93.8 KiB
warehouse	3.6 KiB
warehouse_class	1.2 KiB

Fig. 2: Food Mart Dataset

V. ACCESS CONTROL MODULE

This module authenticates, authorizes and determines the level of privacy for any data share.

A) Role based access control (RBAC): In the given work, there are certain roles which are assigned, given to the administrator, user and manager. The role is assigned with certain access control. The accesses rights are decided based upon read write and update permission. The administrator is equipped with all the access rights, whereas the customer will not have all the rights. Hence role based access control is deciding the access rights of particular individuals.

B) Attribute Based Access Control (ABAC): Attribute based access control is one in which the attributes of different users are analyzed and based on the defined access rights the attribute are refined. Say for example if number of column which customer can access are very less than the access rights which administrator have, hence the number of column which administrator can access are very less than which user are allowed.

VI. RESULT PROCESSOR MODULE

A) K-means clustering

Clustering is an approach to classifying all elements in such a way that every similar element should reside into the single group based on their similarity. Subsequently, it also resides irrelevant elements into another group based on their similarity value and maximum cluster size. . It is one of the simplest unsupervised learning algorithms which simplify the work of mining by classifying the similar elements in a cluster using the k-cancroids parameter. It calculates a distance between each element to evaluate similarity and reside them into a single cluster by comparing with the k-centroid parameter. The Euclidean distance function measures the distance from point A to point B. The equation for this distance between a point P ($y_1, y_2 \ldots y_n$) and point Q ($z_1, z_2 \ldots z_n$) are:

$$D=\sqrt{\sum_{j=1}^{n}(y_j - z_j)^2}$$

A snipping of this work is shown in Fig. 3

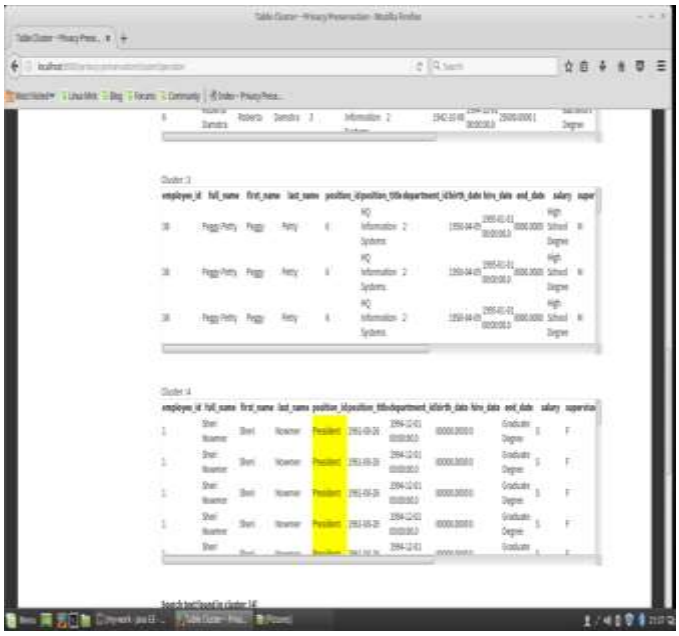


Fig. 3: Using K- means Clustering

B) RC4 Encryption

In cryptography, RC4 (Rivest Cipher 4 also known as ARC4 or ARCFOUR meaning Alleged RC4, see below) is a stream cipher. The main factors in RC4's success over such a wide range of applications have been its speed and simplicity: efficient implementations in both software and hardware were very easy to develop. RC4 is used with K-means clustering to securely send the data. For maintaining the privacy of data can be send in encrypted form. A snipping of this work is shown in Fig. 4

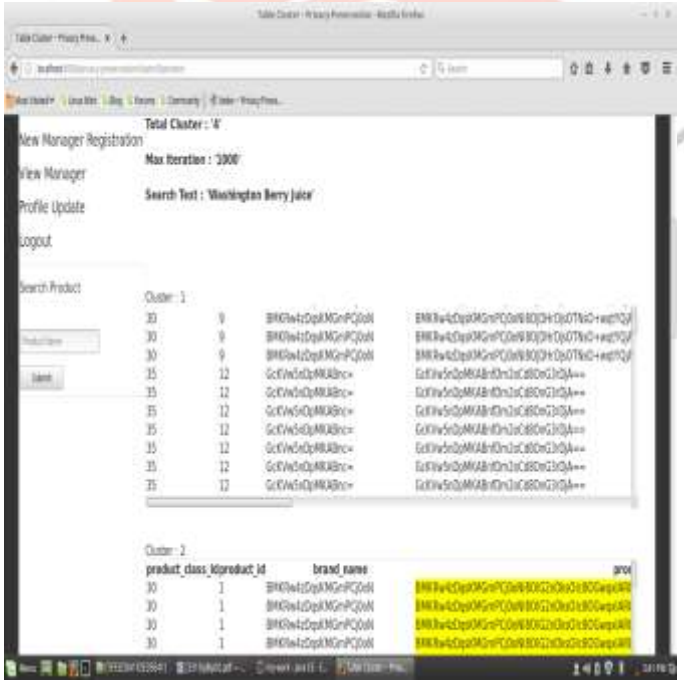


Fig. 4: Encrypted Data using RC4

VII. EXPERIMENT ANALYSIS

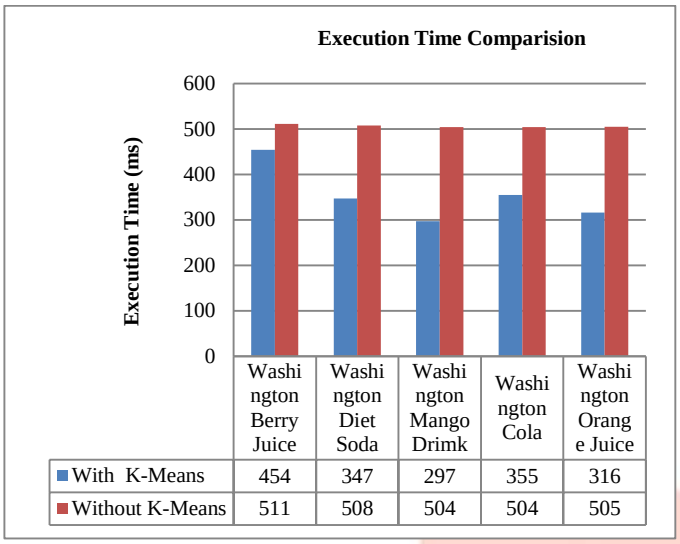
A java based application has been developed to implement the proposed solution. Proposed implementation view has been classified into three modules. Initially, food mart dataset has been exported into MYSQL database server which can be downloaded from Google. Then, we have used access control module. The aim of this module to provide access to the system through mechanism that enforces access control mechanism. After, that authentication process is applied; it would authorize an end user, based on set of rules and also maintain privacy level for the shared data. For this purpose we have used the hybrid access control policies i.e. RBAC and ABAC. After that we have used the K means clustering to group similar elements in one group and dissimilar into other group. After that we have applied the with encryption algorithm to encrypt data.

For that purpose we have used RC4 to encrypt data so that privacy can be achieved and to provide the result to authenticate user. Execution time and recall parameters have been used to measure the performance of proposed solution.

VIII. RESULT ANALYSIS

The complete performance has been evaluated on basis of Recall [Accuracy] and execution time. Initially, we have used five different products to calculate the execution time without K-means clustering. Then, we have to calculate the execution time with K-means clustering with cluster size=4.

Table 1: Comparing Algorithm for Execution Time



Name of Product	K Means with Encryption Time in(ms)	Without K Means Time in(ms)
Washington Berry Juice	454	511
Washington Diet Soda	347	508
Washington Mango Drink	297	504
Washington Cola	355	504
Washington Orange Juice	316	505

Fig. 5 Comparing Algorithm for Execution Time

Fig. 4 shows comparative study of execution time using with K- means and Without K-means clustering algorithm. In this figure the X axis shows the different products and Y axis shows Time in milliseconds. The complete evaluations considered different words as the input and observed the computation time before K-means and After K-means. A reduced computation time has been observed after integration of K-means clustering algorithm. The proposed algorithm is more efficient and secure.

No. of Clusters K	Recall			
	1	2	3	4
4	0.28	0.34	0.28	0.18
5	0.28	0.28	0.28	0.14
6	0.38	0.42	0.32	0.14
7	0.34	0.54	0.36	0.28

Table 2: Comparing Algorithm for Recall vs. No. of Clusters

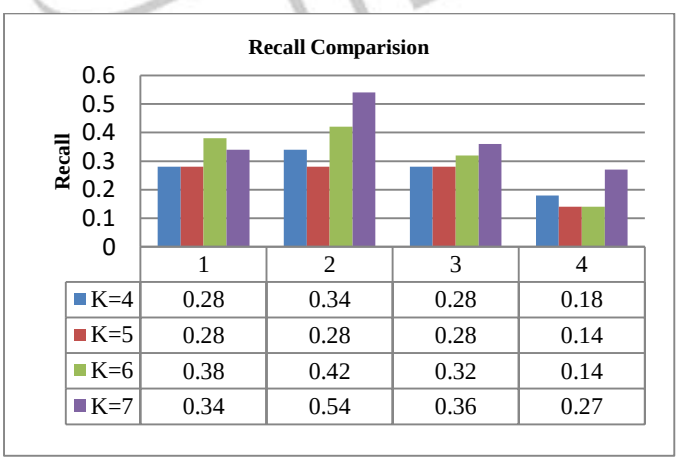


Fig 6: Comparing Algorithm for Recall vs. No. of Clusters

Fig. 5 shows comparative study of algorithm for recall and no. of clusters. We have calculate the different values for different clusters such as K=4, 5, 6, 7.We have observed that when number of clusters increases we have obtain more accuracy as compared to number of clusters is less. Recall has been calculated for different cluster values where minimum is 0.14 and maximum is 0.54 have been recorded.

IX. CONCLUSION

The complete work concludes that proposed work will not only suggest a solution to implement access control mechanism with proposed security model but will help to achieve better performance in large data set. Here a hybrid security model has been proposed based on ABAC and RBAC to manage table permissions and access. A food mart dataset has been considered as the source of information and data schema to implement the proposed solution and evaluate the performance.

The complete solution implements the four different user roles and classifies their responsibility and security according to attributes. Complete solution is evaluated on the basis of measuring the performance in terms of execution time with and without K-means algorithm and recall. A reduced computation time has been observed after integration of K-means clustering algorithm. The complete work concludes that integration of K-means clustering help to improve performance of search operation in comparison with and without K-means approach. We have also observed that when number of clusters increases we have obtain more accuracy as compared to number of clusters is less. Recall has been calculated for different cluster values where minimum is 0.14 and maximum is 0.54 have been recorded. The complete work ends with satisfactory results.

REERENCES

1. Antorweep Chakravorty, Tomasz Wlodarczyk, Chunming Rong "*Privacy Preserving Data Analytics for Smart Homes*" published in IEEE Security and Privacy Workshops, 2013.
2. R. Sandhu, E. Coyne, et. al., "Role-Based Access Control Models," IEEE Computer, vol.29, no.2, pp.38-47, Feb. 2003.
3. 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.
4. R. Bayardo, R. Agrawal, "Data Privacy Through Optimal k- Anonymization," 21th International Conference on Data Enginnering, pp.217-228, Apr. 2005.
5. A Solution For Privacy Protection In MapReduce Quang Tran,Hiroyuki Sato Graduate School of Engineering, The University of Tokyo.
6. S. Meyer, A. Rakotonirainy, "A survey of research on context-aware homes," Australian Computer Society, vol.21, pp.159-168, 2003.
7. G. Drosatos, P. Efraimidis, "Privacy-preserving statistical analysis on ubiquitous health data," 8th International Conference on Trust, Privacy and Security in Digital Business, Springer-Verlag, pp.24-36, 2011.
8. S. Bagüés , A. Zeidler, et. al., "Sentry@Home - Leveraging the Smart Home for Privacy in Pervasive Computing," International Journal of Smart Home, vol.1, no.2, Jul. 2007.
9. L. Sweeney, "Datafly: A system for providing anonymity in medical data," 11th International Conference on Database Security, pp.356–381, 1998.