*International Journal of Research in Advent Technology, Vol.6, No.5, May 2018*
*E-ISSN: 2321-9637*
*Available online at www.ijrat.org*

# A Novel scheme using Unsupervised learning to Improve Privacy in Health care Application

Abhay Raj Dwivedi[1], Akhilanand Mishra[2], Ashutosh Kumar[3], Madhuri S[4]
*Information Science and Engineering[1,2,3,4],SJBIT[VTU][1,2,3,4]*
*Email: abhayrdwivedi@gmail.com[1] ,akhilmishra744@gmail.com[2]*

**Abstract-**Security in the modern days have become very important in daily life due to the huge increase in data. Therefor a lot of clustering techniques have been widely adopted in many real world data analysis, targeted marketing, digital forensics etc. With the huge increase of data in today's big data era, a major way to handle a clustering over large dataset is to out sourcing it to public cloud platform. It is because cloud computing is reliable and offers saving on making infrastructure and also provides a better performance. If such outsourcing is done there are a lot of sensitive information within these datasets like the information about the health of the patients and the medicines or drugs being supplied to the patients.

Here we are proposing a novel scheme using unsupervised learning to improve privacy in health care application. So that no private data is compromised.

**Index Terms-Privacy-preserving ; K-means Clustering ; Cloud Computing**

## 1. INTRODUCTION

Clustering is very important in storing a huge amounts of data on cloud. There are a lot of clustering techniques being used for the storage of big data on cloud. Since clustering and storing the data makes it easier to handle within the cloud. But clustering has many different challenges like the volume of data, variety of the data and the velocity of the data. In order to manage these problems effectively and efficiently the public cloud servers have the major role in order to provide high performance economically. It is necessary because the public cloud servers have an open environment and are maintained by an external third-party. The information of a medical center is equally important to be secure for example Health Insurance Portability and Accountability Act (HIPAA) since these data can be analyzed and can be misused by many pharmaceutical companies.

Here we are using a novel scheme using unsupervised learning under which we propose the use of k-means clustering to accommodate a large amount of data and then equally dividing it into k clusters. Since the data needs to be protected from the third party also so we need to hide our information from the third party and for that we use an user end encryption so that the data is not even recognizable to the third-party. In this context we are using AES (Advanced Encryption Standard) algorithm to generate the public and private keys for the encrypted data. So the people with the keys are only able to modify or read the data that is being stored on cloud and not any TPA (Third Party Association) can view or edit it, they can just do the clustering and storing of the data.

But apart from protecting our privacy there are two more factors in doing k-means clustering over outsourced datasets the first one is the efficiency of the clustering and its accuracy. In other words the k-means clustering in outsourced data must be parallelized which is must for cloud computing environment and its performance on large data. Apart from this the effective cost of the owner should also be minimum. Although there are many MapReduce based K-means clustering techniques proposed but we offer the protection of the outsourced data, neither if the techniques present earlier provides such a feature.

## 2. OUR WORK

In this paper, we are proposing a novel scheme using unsupervised learning to improve privacy in health care application. Through this a largescale data can be outsourced in a very effective way to the public cloud servers, which provides the privacy-preserving to the data objects directly on the ciphertext. In this designing we mainly focus on three different stages encryption, clustering and privacy update. The framework on which we are performing the clustering of the datasets is MapReduce framework which works as the internal layer of the big data architecture.

### 2.1 MapReduce framework

Here in MapReduce the processing of large datasets are done in distributed manner. All the datasets are first mapped within the cloud servers at different clusters and the processing carries on in the different clusters

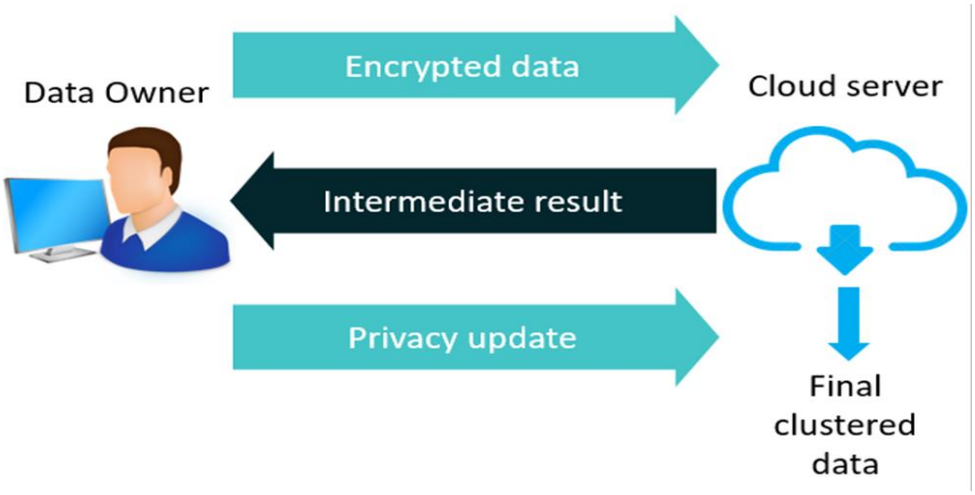



**Fig. 1. System Architecture**

itself so the processing is very fast in such a manner, since all the data are processed at its clusters itself the first letter of the first word only.

## 3. Equation

Here we are considering the data with n data objects which needs to be clustered in k number of clusters. Here each data object is having m elements which is being clustered.

$$\vec{D}_i = [r_i d_{i1}, r_i d_{i2}, \cdots, r_i d_{im}, r_i, \alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{im-1}]$$
$$\vec{D}'_i = [d_{i1}, d_{i2}, \cdots, d_{im}, r_i, \alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{im-1}]$$

where *ri, αi1, αi2, ⋯, αi(m−1) ∈ Zp* are all random numbers selected by users. The owner also selects the initial number of clusters K.

$$\vec{C}_k = [c_{k1}, c_{k2}, \cdots, c_{km}, -\frac{1}{2}\sum_{j=1}^{m} c_{kj}^2, \beta_1, \beta_2, \cdots, \beta_{m-1}]$$

where *β1,β2,⋯ ,βm−1 ∈ Zp* are random numbers selected by the owner.

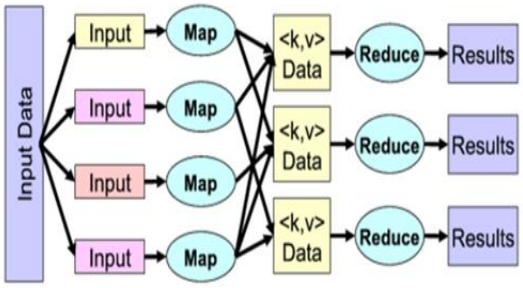



**Fig 2. MapReduce framework**

### 3.1 Key generation:

The generation of the secret key is done by the use of AES algorithm here we use 128 bits key generation which will keep the iteration going until 10 iterations are complete after completing 10 iterations then the secret key is generated since we are using a symmetric algorithm here the generation of the key is symmetric i.e the public key and the private key have equal importance and the organizations having these keys can edit or read the data. While the TPA is not able to see the data since it has only the permission for performing the clustering and not editing or reading.

This algorithm uses left shift and reduce technique to perform the key generation. The data is stored in the form of matrix and first the matrix is shifted and then key is generated the key generated here is of 128 bits.

The architecture consists of 3 main processes the encrypted data goes to the server for clustering then the cloud server performs the clustering then the clustered intermediate result data is sent back to the user for verification then if the user wants to edit or read or add the new information to the data then the user does it and sends the privacy update wether there needs to be more clustering or not. Or if there is more data needs to be added to the clustered datasets.

### 3.2 Related work

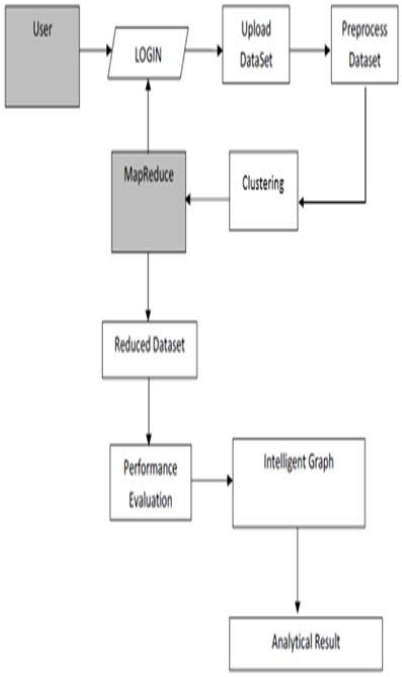
There are a number of schemes in accordance to outsourced data clustering in privacy preserving way. In distance preserving data transformation schemes are adopted to protect the privacy of dataset, while keeping the distance comparison property for clustering purpose. These schemes are very efficient and even achieve the same computational cost compared to the original clustering algorithm. This is because data perturbation based encryption makes the ciphertext have the same size as the original data, and uses the

same clustering operations in the original clustering algorithms.

## 4 PROCESS FLOW

Here the figure below denotes how the flow of data goes within the complete system. The users first login and then upload the data the data uploaded is encrypted and the keys are generated these keys stays with the owner. This data is termed as preprocessed datasets, then this data is sent for further Map and Reduce work after the MapReduce processing on the datasets are done then the reduced dataset are further sent for performance evaluation and then the results are generated based on the performance evaluation done. The intelligent graph plotted with the information provided by the analysis data makes the owner know about the cost and effectiveness as well as the ciphertext strength and the amount of data encrypted and the amount of data that is still not encrypted it is an end to end analysis of the datasets.

### 4.1 Flow chart




## 5 CONCLUSION

Here in our system, we proposed a privacy-preserving MapReduce based K-means clustering scheme in cloud computing. Thanks to our encryption design based on the AES hard problem, our scheme achieves clustering speed and accuracy that are comparable to the K-means clustering without privacy-protection. Considering the support of large-scale dataset, we securely integrated MapReduce framework into our design, and make it extremely suitable for parallelized processing in cloud computing environment. In addition, the privacy preserving Euclidean distance comparison component proposed in our design can also be used as an independent tool for distance based applications. We provide thorough analysis to show the security and efficiency of our scheme. Our prototype implementation over 5 million data objects demonstrates that our scheme is efficient, scalable, and accurate for K-means clustering over large-scale data.

## REFERENCES

[1]. European Network and Information Security Agency. Cloud computing security risk assessment. https://www.enisa.europa.eu/activities/riskmanagement/files/deliverables/cloud-computing-risk-assessment.

[2]. Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, and Albert-L´aszl´o Barabasi. Predicting individual disease risk based on medical history. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM ’08, pages 769– 778, Napa Valley, California, USA, 2008.

[3]. U.S. Dept. of Health & Human Services. Standards for privacy of individually identifiable health information, final rule, 45 cfr, pt 160– 164. http://www.hhs.gov/sites/default/files/introduction.pdf, 2002.

[4]. Jaideep Vaidya and Chris Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’03, pages 206–215, New York, NY, USA, 2003. ACM.

[5]. Geetha Jagannathan and Rebecca N. Wright. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD ’05, pages 593–599, New York, NY, USA, 2005. ACM.

[6]. Paul Bunn and Rafail Ostrovsky. Secure two-party k-means clustering. In Proceedings of the 14th ACM Conference on Computer and Communications Security, CCS ’07, pages 486–497, New York, NY, USA, 2007. ACM.

[7]. Mahir Can Doganay, Thomas B. Pedersen, Y¨ucel Saygin, Erkay Savas¸, and Albert Levi. Distributed privacy preserving k-means clustering with additive secret sharing. In

Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, PAIS '08, pages 3–11, New York, NY, USA, 2008. ACM.

[8]. Jun Sakuma and Shigenobu Kobayashi. Large-scale k-means clustering with user-centric privacy-preservation. Knowledge and Information Systems, 25(2):253–279, 2009.

[9]. Xun Yi and Yanchun Zhang. Equally contributory privacy-preserving k-means clustering over vertically partitioned data. Inf. Syst., 38(1):97– 107, March 2013. 00000
Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. SIGMOD Rec., 29(2):439–450, May 2000.

[10]. Stanley R. M. Oliveira and Osmar R. Zaane. Privacy preserving clustering by data transformation. In Brazilian Symposiumon Databases, SBBD, Manaus, Amazonas, Brazil, 2003.

[11]. Kun Liu, Chris Giannella, and Hillol Kargupta. An attacker's view of distance preserving maps for privacy preserving data mining. In Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases, PKDD'06, pages 297–308, Berlin, Heidelberg, 2006. Springer-Verlag.

[12]. H. Kargupta, S. Datta, Q. Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties of random data perturbation techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, pages 99–106, Nov 2003.

[13]. Dongxi Liu, Elisa Bertino, and Xun Yi. Privacy of outsourced k-means clustering. In Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security, ASIA CCS '14, pages 123– 134, New York, NY, USA, 2014. ACM.

[14]. Yongge Wang. Notes on two fully homomorphic encryption schemes without bootstrapping. Cryptology ePrint Archive, Report 2015/519, 2015.