Secure Data Mining Using Knn Algorithm In Cloud Computing

Dr.T.S.Baskaran[1], E.Eniyamudhu[2]
[1, 2] Dept of Computer Science
[1, 2] A.Veeriya Vandayar.Memorial Sri Pushpumcollege

Abstract- *A large quantity of compound and enveloping Digital data is being generated which is growing at an Storage space systems are not able to handle Big Data and also Analyzing the Big Data become a confront and therefore it cannot be resolve the difficulty of handling, storage space s and analyzing the Big . Cloud Computing is the best answer available to the trouble of Big Data storage space and its analyses but having said that, there is Always a possible risk to the security of Big Data storage space in* one of the major issues while storing the Big Data in a Cloud *environment. Data Mining based assault, a key danger to the data, allows an adversary or an unauthorized user to infer valuable and sensitive information by analyzing the results generated from computation performed on the raw data. This thesis recommend a secure knn data mining advance haughty the data to be scattered among different hosts preserving the The advance is able to continue the precision and authority of the existing k-means to generate the In machine learning, people often confused with k-means (k-means clustering) and KNN (k-Nearest Neighbors). K-means is an unsupervised learning algorithm used for clustering problem whereas KNN is a supervised learning algorithm used for classification and regression problem.*

Keywords- Storage Space, Cloud, Big Data, KNN (k-Nearest Neighbours), Security, Encryption.

I. INTRODUCTION

1.1Cloud computing:

Cloud computing is the result of growth and approval of existing technologies and paradigms. The goal of cloud computing is to allow users to take benefit from all of these technology, without the call for for deep facts about or knowledge with each one of them. The cloud aims to cut costs, and help the users heart on their core business instead of being impeded by IT obstacle.

The major enable technology for cloud computing is virtualization. By minimizing user involvement, mechanization speeds up the procedure, reduce labor costs and reduce the likelihood of human errors.

Users characteristically facade difficult business problems. Cloud computing adopts concepts from Service-oriented Architecture (SOA) that can help the user break these problems into military that can be integrated to provide a answer. Cloud computing provide all of its resources as services, and makes use of the entrenched standards and best practices gain in the area of SOA to allow global and easy access to cloud services in a homogeneous way.

- Grid computing — "A form of distributed and parallel computing, whereby a 'super and virtual computer' is composed of a cluster of networked, loosely coupled computers acting in concert to perform very large tasks."
- Mainframe computer — Powerful computers used mainly by large organizations for critical applications, typically bulk data processing such as: census; industry and consumer statistics; police and secret intelligence services; enterprise resource planning; and financial transaction processing.
- Utility computing — The "packaging of computing resources, such as computation and storage, as a metered service similar to a traditional public utility, such as electricity."
- Peer-to-peer — A distributed architecture without the need for central coordination. Participants are both suppliers and consumers of resources (in contrast to the traditional client–server model).

These journalists present an come within reach of to excavation the data securely using kNN algorithm from the cloud even in the attendance of adversaries. This proposed approach prevents any intermediate dataleakage in the process of computation while maintaining the correctness and validity of the data.

II. EXISTING APPROACH

Notations: Ci represents the combined clustering centers which is the sum of Host A and Host B's share i.e. HA and HBrespectively where Ci= HA+ HB.

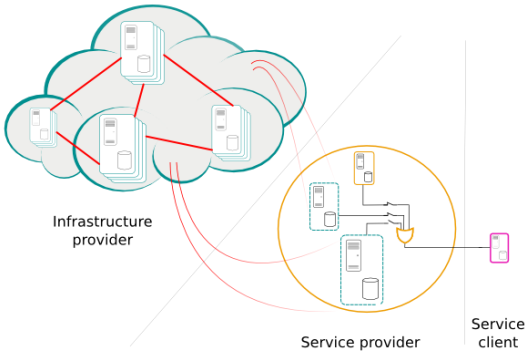


Fig 1: Cloud Computing

Input:

1) Database DA and DB belonging to Host A and Host B respectively having n data objects.
2) 'k' which is the total number of clusters.

Output:

The k cluster which is the combination of DA andDB or D.

1) Each party performs Data Normalization on local data.
2) Host A and Host B select their respective k cluster centers H1A, H2A,……., HK.
3) Calculate or perform local k-means for Host A and Host B.
4) Save the cluster centers Hj A,i, HjB,i .
5) Perform the secure cluster updation and reassign the data objects to their closest clusters locally.
6) Save Hj,A,i+1, HjB,i+1 . if the difference between the previous cluster center and the current one is less than or equal to threshold value then stop the iteration else repeat step 4 onwards.

Figure 1. Overview of the Cloud Approach

III. DETAILED APPROACH

The proposed approach uses the public key cryptosystems where M is the message or the plain text which is to beencrypted. The system can be divided into 3 parts (K,E,D):

- A pair of public and private key (lk,pk) is generated.
- A ciphertext or encrypted message c=Elk(m,r) is obtained where m € M and r is a random value.
- Decryption Dpk(c)=m is used to obtain plain text again.

DISADVANTAGE:

- There is always a potential risk to the security of Big Data storage in Cloud Computing.
- Data privacy is the one of the major issue

The storage on cloud can mine the data and retrieve large amount of confidential data.

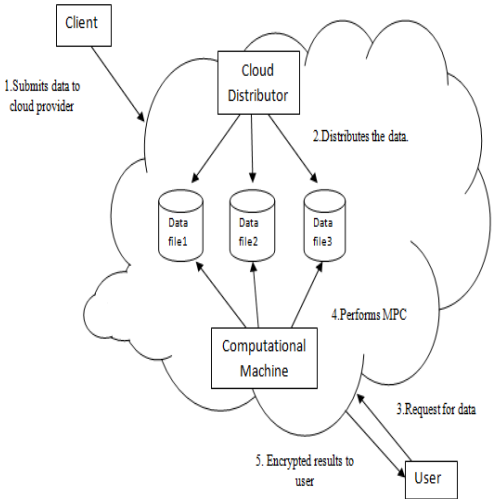


Fig 2: Existing Architecture

IV. PROPOSED APPROACH

k-NN is a type of instance-based learning, or lazy learning, where the purpose is only approximated locally and all calculation is delayed until classification. The k-NN algorithm is among the simplest of all mechanism learning algorithms. together for classification and regression, a useful procedure can be used to give weight to the contributions of the neighbours, so that the nearer neighbours donate more to the usual than the more distant ones. For example, a widespread weighting scheme consists in giving each neighbour a weight of 1/d, where d is the distance to the neighbour.

The proposed approach can further be extended by adding a digital signature or hashing technique to authenticate the third party so as to prevent an adversary from posing as the third party to host's.This approach assumes that the data is not stored in a centralized location but is distributed to various hosts.

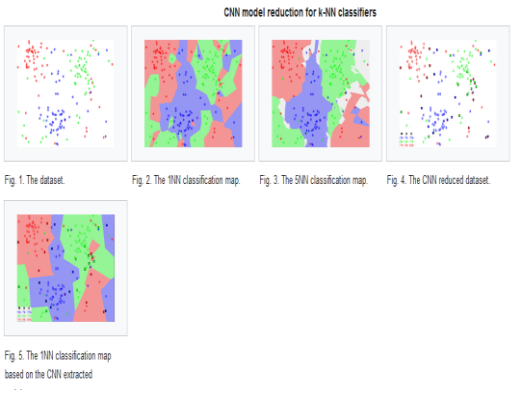


Fig 3: KNN Classification

ALGORITHM

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour.

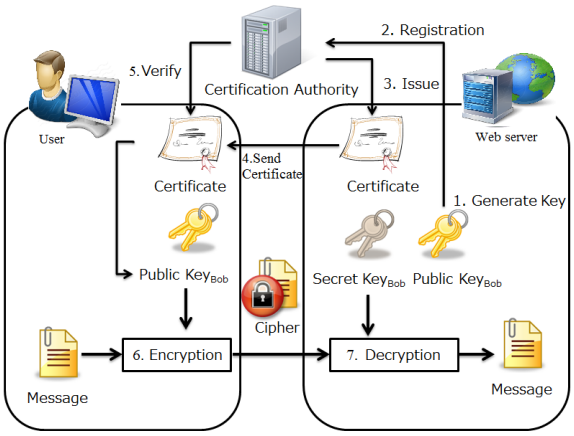
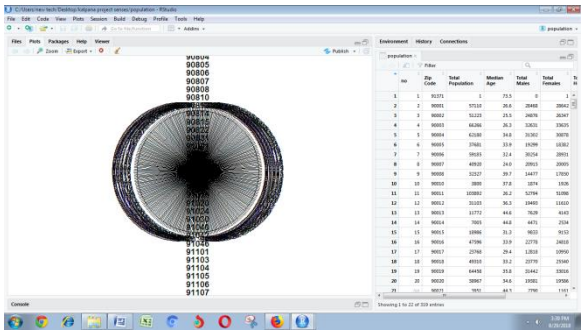


Fig 4: Proposed Architecture

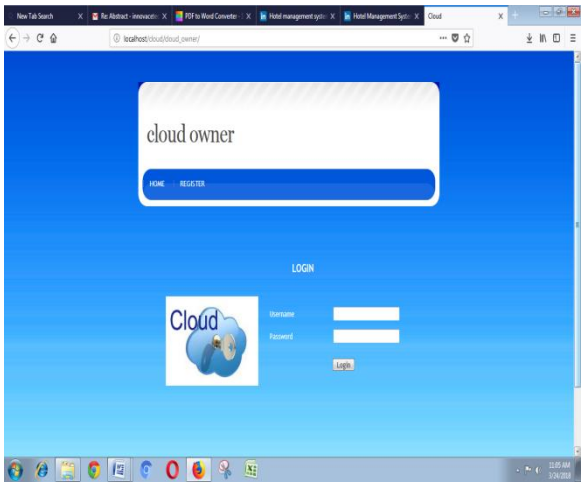
V. SYSTEM IMPLEMENTATION



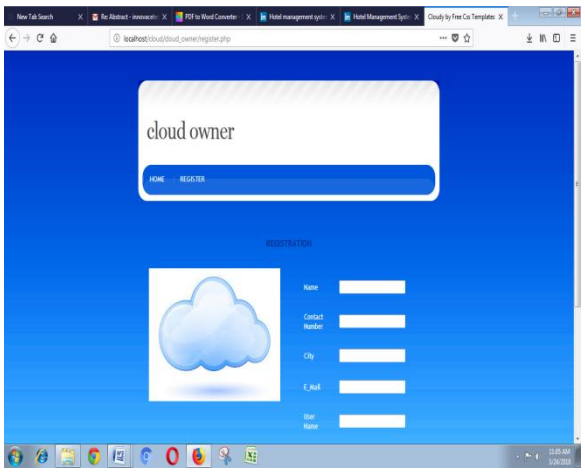
In this System We Propose R tool The above figures show the correctness of the proposed algorithm. It can be seen that the final classification

Thus, it is proved that the algorithm maintains the correctness and validity of the final result and thus canbe applied to all situations where a single party knn can be used.

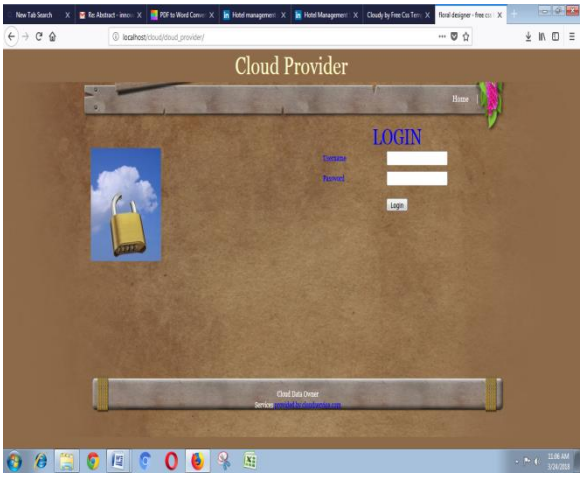
CLOUD COMPUTING INTERFACES:



Cloud Owner



Cloud File Acess



Cloud Provider

VI. CONCLUSION

Security and privacy is the main problem about the clients as well as the provider of cloud services as a lot of confidential and sensitive data is stored in cloud which can provide valuable information to an attacker. K-means is an unsupervised learning algorithm used for clustering problem whereas KNN is a supervised learning algorithm used for classification and regression problem. This document propose a method to solve the privacy issues of the cloud. It assume that the user data is distributed on two hosts and performs a combined knn Classification using the encryption system for security purpose so as to prevent any interpretation of intermediate results by an attacker. This system overcome the existing approach and provide security.

REFERENCES

[1] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOM, pp. 829-837, Apr, 2011.

[2] L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Towards a Cloud Definition," ACM SIGCOMM Comput. Commun. Rev., vol. 39, no. 1, pp. 50-55, 2009.

[3] N. Cao, S. Yu, Z. Yang, W. Lou, and Y. Hou, "LT Codes-Based Secure and Reliable Cloud Storage Service," Proc. IEEE INFOCOM, pp. 693-701, 2012.

[4] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptograpy and Data Security, Jan. 2010.

[5] A. Singhal, "Modern Information Retrieval: A Brief Overview," IEEE Data Eng. Bull., vol. 24, no. 4, pp. 35-43, Mar. 2001.

[6] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishing, May 1999.

[7] D. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2000.

[8] E.-J. Goh, "Secure Indexes," Cryptology ePrint Archive, http:// eprint.iacr.org/2003/216. 2003.

[9] Y.-C. Chang and M. Mitzenmacher, "Privacy Preserving Keyword Searches on Remote Encrypted Data," Proc. Third Int'l Conf. Applied Cryptography and Network Security, 2005.

[10] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," Proc. 13th ACM Conf. Computer and Comm. Security (CCS '06), 2006.

[11] D. Boneh, G.D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public Key Encryption with Keyword Search," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2004.

[12] M. Bellare, A. Boldyreva, and A. ONeill, "Deterministic and Efficiently Searchable Encryption," Proc. 27th Ann. Int'l Cryptology Conf. Advances in Cryptology (CRYPTO '07), 2007.

[13] M. Abdalla, M. Bellare, D. Catalano, E. Kiltz, T. Kohno, T. Lange, J. Malone-Lee, G. Neven, P. Paillier, and H. Shi, "Searchable Encryption Revisited: Consistency Properties, Relation to Anonymous Ibe, and Extensions," J. Cryptology, vol. 21, no. 3, pp. 350- 391, 2008.

[14] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy Keyword Search Over Encrypted Data in Cloud Computing," Proc. IEEE INFOCOM, Mar. 2010.

[15] D. Boneh, E. Kushilevitz, R. Ostrovsky, and W.E.S. III, "Public Key Encryption That Allows PIR Queries," Proc. 27th Ann. Int'l Cryptology Conf. Advances in Cryptology (CRYPTO '07), 2007.

[16] P. Golle, J. Staddon, and B. Waters, "Secure Conjunctive Keyword Search over Encrypted Data," Proc. Applied Cryptography and Network Security, pp. 31-45, 2004.

[17] L. Ballard, S. Kamara, and F. Monrose, "Achieving Efficient Conjunctive Keyword Searches over Encrypted Data," Proc. Seventh Int'l Conf. Information and Comm. Security (ICICS '05), 2005.

[18] D. Boneh and B. Waters, "Conjunctive, Subset, and Range Queries on Encrypted Data," Proc. Fourth Conf. Theory Cryptography (TCC), pp. 535-554, 2007.

[19] R. Brinkman, "Searching in Encrypted Data," PhD thesis, Univ. of Twente, 2007.

[20] Y. Hwang and P. Lee, "Public Key Encryption with Conjunctive Keyword Search and Its Extension to a Multi-User System," Pairing, vol. 4575, pp. 2-22, 2007.