© 2016 IJSRSET | Volume 2 | Issue 3 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Data Mining Over Encrypted Data on Cloud

Suganya. R, M. Nizar Ahmed

Computer Science Department, PPG Institute of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Data Mining has wide applications in many areas such as medicine, scientific, banking, research and among government agencies. For the past decade, due to the rise of various privacy issues, many theoretical and practical solutions to the classification problem have been proposed under different security models. Classification is one of the commonly used tasks in data mining applications However, with the recent popularity of cloud computing, users now have the opportunity to outsource their data, in encrypted form, as well as the data mining tasks to the cloud. Since the data on the cloud is in encrypted form, existing privacy-preserving classification techniques are not applicable. In this paper, we focus on solving the classification problem over encrypted data. In particular, we propose a secure k-NN classifier over encrypted data in the cloud. The proposed protocol protects the confidentiality of data, privacy of user's input query, and hides the data access patterns. To the best of our knowledge, our work is the first to develop a secure k-NN classifier over encrypted data under the semi-honest model. Also, we empirically analyze the efficiency of our proposed protocol using a real-world dataset under different parameter settings. The proposed system mainly focuses on information security in insurance company. They can encrypt the customer information and stored it in database. When data are encrypted, any data mining tasks becomes very challenging before decrypting data. Classification can apply to the customer records. This protects the customers' sensitive information.

Keywords : Data Mining, Encrypted Database, Security, K-NNclassifier

I. INTRODUCTION

Data mining techniques and applications are very much needed in the cloud computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining. "Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users."

As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way. The main effects of data mining tools being delivered by the Cloud are: The customer only pays for the data mining tools that he needs that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive; The customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing. Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments.

Cloud Computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users." The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.

The relationship between data mining and cloud is worth to discuss. Cloud providers use data mining to provide clients a better service. If clients are unaware of the information being collected, ethical issues like privacy and individuality are violated. This can be a serious data privacy issue if the cloud providers misuse the information. Again attackers outside cloud providers having unauthorized access to the cloud, also have the opportunity to mine cloud data. In both cases, attackers can use cheap and raw computing power provided by cloud computing to mine data and thus acquire useful information from data. As cloud is a massive source of centralized data, data mining gives attackers a great advantage in extracting valuable information and thus violating clients' data privacy.

II. METHODS AND MATERIAL

1. Literature Survey

A. Fully Homomorphic Encryption Using Ideal Lattices

This design proposed a fully homomorphism encryption scheme i.e., a scheme that allows one to evaluate circuits over encrypted data without being able to decrypt. This solution comes in three steps. First, it provides a general result – that, to construct an encryption scheme that permits evaluation of arbitrary circuits, it suffices to construct an encryption scheme that can evaluate (slightly augmented versions of) its own decryption circuit; it call a scheme that can evaluate its (augmented) decryption circuit boots trappable.

Next, it describes a public key encryption scheme using ideal lattices that is almost boots trappable. Lattice-based cryptosystems typically have decryption algorithms with low circuit complexity, often dominated by an inner product computation that is in NC1. Also, ideal lattices provide both additive and multiplicative homeomorphisms (modulo a public-key ideal in a polynomial ring that is represented as a lattice), as needed to evaluate general circuits.

Unfortunately, the initial scheme is not quite boots trappable i.e., the depth that the scheme can correctly evaluate can be logarithmic in the lattice dimension, just like the depth of the decryption circuit, but the latter is greater than the former. In the final step, it shows how to modify the scheme to reduce the depth of the decryption circuit, and thereby obtain a bootstrappable encryption scheme, without reducing the depth that the scheme can evaluate. Abstractly, it accomplish this by enabling the encryptor to start the decryption process, leaving less work for the decrypter, much like the server leaves less work for the decrypter in a server-aided cryptosystem.

B. Privacy Preserving Naïve Bayes Classifier for Horizontally Partitioned Data

The problem of secure distributed classification is an important one. In many situations, data is split between multiple organizations. These organizations may want to utilize all of the data to create more accurate predictive models while revealing neither their training data / databases nor the instances to be classified. The Naive Bayes Classifier is a simple but efficient baseline classifier. In this paper, authors present a privacy preserving Naive Bayes Classifier for horizontally partitioned data.

The Naive Bayes classifier is a simple but efficient baseline classifier. It is the de facto classifier used for text classification. Naive Bayes is based on a bayesian formulation of the classification problem which uses the simplifying assumption of attribute independence. It is simple to implement and use while giving surprisingly good results. Thus, preliminary evaluation is carried out using the Naive Bayes classifier to serve both as a baseline and to decide whether more sophisticated solutions are required. The problem of secure distributed classification is an important one. The goal is to have a simple, efficient and privacy-preserving classifier. The ideal would be for all parties to decide on a model. Jointly select/discover the appropriate parameters for the model and then use the model locally as and when necessary.

In this work that data is horizontally partitioned. This means that many parties collect the same set of information about different entities. Parties want to improve classification accuracy as much as possible by leveraging other parties' data. They do not want to reveal their own instances or the instance to be classified. Thus, what we have is collaboration for their own advantage. One way to solve this is to decide on a model. The model parameters are generated jointly from the local data. Classification is performed individually without involving the other parties. Thus, the parties decide on sharing the model, but not the training set nor the instance to be classified. This is quite realistic. For example, consider banks which decide to leverage all data to identify fraudulent credit card usage, or insurance companies which jointly try to identify high-risk customers. In this paper, authors use / extend several existing cryptographic techniques to create a privacy preserving Naive Bayes Classifier for horizontally partitioned data.

C. Privacy Preserving Mining of Association Rules

Authors present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. While it is feasible to recover association rules and preserve privacy using a straightforward uniform" randomization, the discovered rules can unfortunately be exploited to find privacy breaches. It analyzes the nature of privacy breaches and proposes a class of randomization operators that are much more effective than uniform randomization in limiting the breaches. It derives formulae for an unbiased support estimator and its variance, which allow recovering itemset supports from randomized datasets, and showing how to incorporate these formulae into mining algorithms. Finally, it presents experimental results that validate the algorithm by applying it on real datasets.

Authors continue the investigation of the use of randomization in developing privacy-preserving data mining techniques, and extend this line of inquiry along two dimensions:

i. Categorical data instead of numerical data, and
ii. Association rule mining instead of classification.

D. Secure kNN Computation on Encrypted Databases

Service providers like Google and Amazon are moving into the SaaS (Software as a Service) business. They turn their huge infrastructure into a cloud-computing environment and aggressively recruit businesses to run applications on their platforms. To enforce security and privacy on such a service model, we need to protect the data running on the platform. Unfortunately, traditional encryption methods that aim at providing unbreakable protection are often not adequate because they do not support the execution of applications such as database queries on the encrypted data. In this paper authors discuss the general problem of secure computation on an encrypted database and propose a SCONEDB (Secure Computation ON an Encrypted Database) model, which captures the execution and security requirements. As a case study, they focus on the problem of k-nearest neighbor (kNN) computation on an encrypted database. Authors develop a new asymmetric scalar-product-preserving encryption (ASPE) that preserves a special type of scalar product. It uses APSE to construct two secure schemes that support kNN computation on encrypted data; each of these schemes is shown to resist practical attacks of a different background knowledge level, at a different overhead cost. Extensive performance studies are carried out to evaluate the overhead and the efficiency of the schemes.

2. Existing System

Existing work on privacy-preserving data mining (PPDM) (either perturbation or secure multi-party computation (SMC) based approach) cannot solve the DMED problem. Perturbed data do not possess semantic security, so data perturbation techniques cannot be used to encrypt highly sensitive data. Also the perturbed data do not produce very accurate data mining results. Secure multi-party computation based approach assumes data are distributed and not encrypted at each participating party.

We claim that the PPkNN problem cannot be solved using the data distribution techniques since the data in our case is encrypted and not distributed in plaintext among multiple parties. For the same reasons, we also do not consider secure k-NN methods in which the data are distributed between two parties.

PPkNN is a more complex problem than the execution of simple kNN queries over encrypted data. For one, the intermediate k-nearest neighbors in the classification process should not be disclosed to the cloud or any users. We emphasize that the recent method in reveals the k-nearest neighbors to the user. Second, even if we know the k-nearest neighbors, it is still very difficult to find the majority class label among these neighbors since they are encrypted at the first place to prevent the cloud from learning sensitive information. Third, the existing work did not address the access pattern issue which is a crucial privacy requirement from the user's perspective.

E. Disadvantages

1. Existing techniques are very expensive
2. They do not produce accurate data mining results due to the addition of statistical noises to the data.
3. Not secured.

3. Problem Definition

Applying data mining technique over encrypted data in the cloud is challenging one. The user privacy should not reveal to the intermediate persons. Introducing the third party auditor provides less security. Hence the system should provide the accurate data mining results.

F. Proposed Design

We proposed novel methods to effectively solve the DMED problem assuming that the encrypted data are outsourced to a cloud. Specifically, we focus on the classification problem since it is one of the most common data mining tasks. Because each classification technique has their own advantage, to be concrete, this paper concentrates on executing the k-nearest neighbor classification method over encrypted data in the cloud computing environment. The proposed system can be implementing in any of the application.

G. Advantages

- ✓ New security primitives and solutions are introduced
- ✓ It protects the confidentiality of data and hides the data access patterns
- ✓ It is more Efiicient
- ✓ It improves the performance
- ✓ There is no use of any third party auditor

H. Modules

a) Cloud Server:

In this module admin can collect the customer details and stored it in the cloud server. The customer details include customer name, customer ID, age, address, city, marital status, yearly income, type of policy etc.,. Admin can add the type of policy and policy details in the cloud. Admin can maintain all customer details, employee details and other policy details in cloud storage.

b) Encrypting Data:

The details maintaining in the cloud are needed to encrypt in order to protect the data from data leakage. This is done by admin. After encrypting data, no one view the contents of data without decrypting it. For the purpose of data encryption, AES algorithm can be used.

c) Customer Module:

Customer can view the records from the cloud. They can update the records at every time of premium. Customer should login to the system with the respective user id and password. Before login to the system, customer should register with their personal details. Customer can view the type of policies and policy details.

d) Data Miner Module:

Data miner may be the employees in the insurance company. Employees may need some records in order for verification or updating records. If they need decrypted data, they need to send request to admin. After verification of requester details, admin can send the decrypted file or decryption key to the users.

e) Classifying Encrypted Data:

In this module classification method is used in order to classify the encrypted data. Admin can retrieve the records based on some patterns. For example, admin can retrieve the records according to the type of policy, or month wise premium etc. For classifying these data's k-nn classifier can be used.

f) Generating Report:

In this module admin can generate the report based on the retrieval records for ease verification. This report hides the user data access patterns and protects the confidentiality of data.

g) Clustering Results:

In this module, admin can cluster the results based on some criteria for ease verification of records. Admin can view the policy holders based on the selected policy details.

And he can also generate the report based on the date. So the admin can verify the records in month wise basis.

4. System Design

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture.

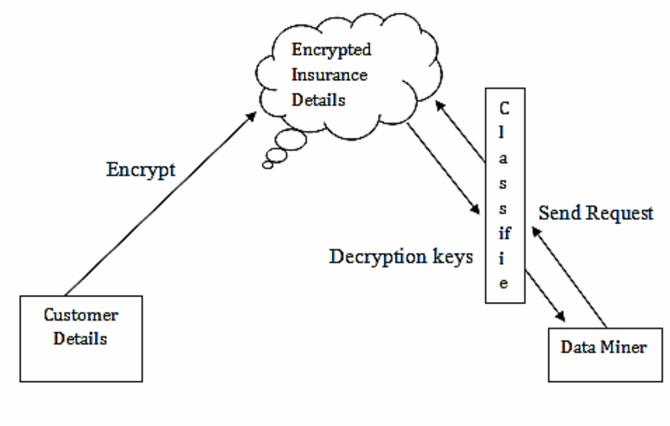


Figure 1. Architecture Diagram

I. File Design

In computing, a file system (or file system) is used to control how data is stored and retrieved. Without a file system, information placed in a storage area would be one large body of data with no way to tell where one piece of information stops and the next begins. By separating the data into individual pieces, and giving each piece a name, the information is easily separated and identified. Taking its name from the way paper-based information systems are named, each group of data is called a "file". The structure and logic rules used to manage the groups of information and their names are called a "file system".

There are many different kinds of file systems. Each one has different structure and logic, properties of speed, flexibility, security, size and more. Some file systems have been designed to be used for specific applications. For example, the ISO 9660 file system is designed specifically for optical discs.

File systems can be used on many different kinds of storage devices. Each storage device uses a different kind of media. The most common storage device in use today is a hard drive whose media is a disc that has been coated with a magnetic film. The film has ones and zeros 'written' on it sending electrical pulses to a magnetic "read-write" head. Other media that are used are magnetic tape, optical disc, and flash memory. In some cases, the computer's main memory (RAM) is used to create a temporary file system for short term use.

J. Input Design

In input design stage, which is the part of the system design stage the system analyst has to decide what inputs are required for the system and prepare input format to give input to the system according to the requirement. Considering the input to the front end from the user we use the user-friendly visual basic software so that the user can easily enter the data.

K. Output Design

Intelligent output design will improve systems relationships with the user and help in decision making. Outputs are also used to provide a permanent hardcopy of the results for latter consultations. The most important reason, which tempts the user to go for a new system is the output. The output generated by the system is often regarded as the criterion for evaluating the usefulness for the system. Here the output requirements use to be predetermined before going to the actual system design. The output design is based on the following:

- Determining the various outputs to be presented to the user.
- Differentiating between inputs to be displayed and those to be printed.
- The format for the presentation of the outputs.

L. Data Flow Diagram

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

Implementation is the final and important phase, the most critical stage in achieving a successful new system and in giving the users confidence. That the new system will work be effective. The system can be implemented only after through testing is done and if it found to working according to the specification. This method also offers the greatest security since the old system can take over if the errors are found or inability to handle certain type of transactions while using the new system.

M. DataBase Design

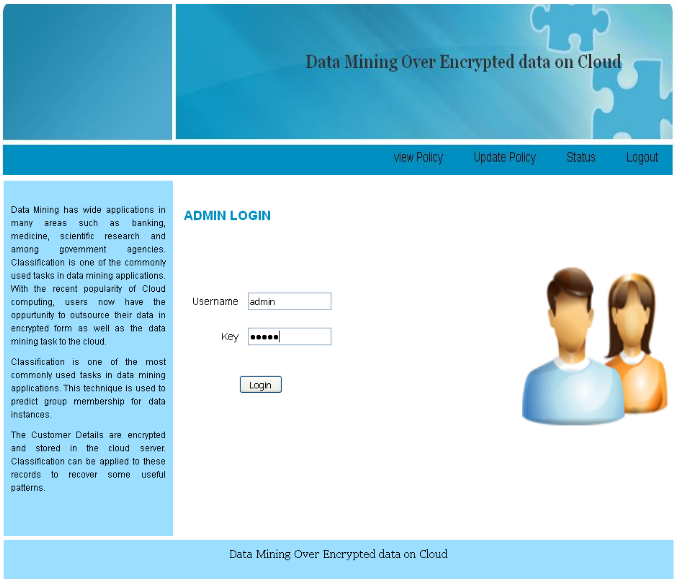
Database design is the process of producing a detailed data model of a database. This logical data model contains all the needed logical and physical design choices and physical storage parameters needed to generate a design in a data definition language, which can then be used to create a database. A fully attributed data model contains detailed attributes for each entity.

The term database design can be used to describe many different parts of the design of an overall database system. Principally, and most correctly, it can be thought of as the logical design of the base data structures used to store the data. In the relational model these are the tables and views. In an object database the entities and relationships map directly to object classes and named relationships. However, the term database design could also be used to apply to the overall process of designing, not just the base data structures, but also the forms and queries used as part of the overall database application within the database management system (DBMS).

The process of doing database design generally consists of a number of steps which will be carried out by the database designer. Usually, the designer must:

- Determine the relationships between the different data elements.
- Superimpose a logical structure upon the data on the basis of these relationships

III. RESULTS AND DISCUSSION



Data Mining has wide applications in many areas such as banking, medicine, scientific research and among government agencies. Classification is one of the commonly used tasks in data mining applications. With the recent popularity of Cloud computing, users now have the oppurtunity to outsource their data in encrypted form as well as the data mining task to the cloud.

Classification is one of the most commonly used tasks in data mining applications. This technique is used to predict group membership for data instances.

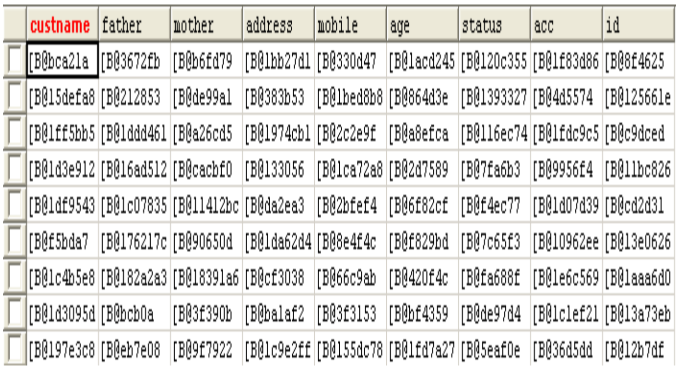
The Customer Details are encrypted and stored in the cloud server. Classification can be applied to these records to recover some useful patterns.

Username admin

Key ●●●●●

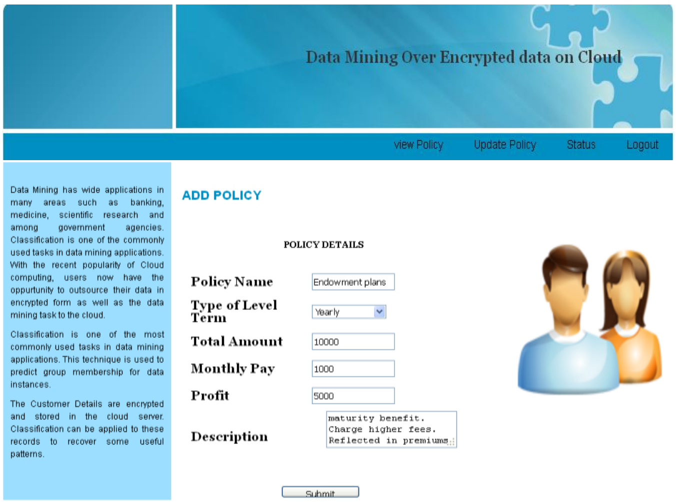
Data Mining Over Encrypted data on Cloud

Figure 2. Admin Login



	custname	father	mother	address	mobile	age	status	acc	id
□	[B@bca21a	[B@3672fb	[B@b6fd79	[B@1bb27d1	[B@330d47	[B@1acd245	[B@120c355	[B@1f83d86	[B@8f4625
□	[B@15defa8	[B@212853	[B@de99a1	[B@383b53	[B@1bed8b8	[B@864d3e	[B@1393327	[B@4d5574	[B@125661e
□	[B@1ff5bb5	[B@1ddd461	[B@a26cd5	[B@1974cb1	[B@2c2e9f	[B@a8efca	[B@116ec74	[B@1fdc9c5	[B@c9dced
□	[B@1d3e912	[B@16ad512	[B@cacbf0	[B@133056	[B@1ca72a8	[B@2d7589	[B@7fa6b3	[B@9956f4	[B@11bc826
□	[B@1df9543	[B@1c07835	[B@11412bc	[B@da2ea3	[B@2bfef4	[B@6f82cf	[B@f4ec77	[B@1d07d39	[B@cd2d31
□	[B@f5bda7	[B@176217c	[B@90650d	[B@1da62d4	[B@8e4f4c	[B@f829bd	[B@7c65f3	[B@10962ee	[B@13e0626
□	[B@1c4b5e8	[B@182a2a3	[B@18391a6	[B@cf3038	[B@66c9ab	[B@420f4c	[B@fa688f	[B@1e6c569	[B@1aaa6d0
□	[B@1d3095d	[B@bcb0a	[B@3f390b	[B@ba1af2	[B@3f3153	[B@bf4359	[B@de97d4	[B@1c1ef21	[B@13a73eb
□	[B@197e3c8	[B@eb7e08	[B@9f7922	[B@1c9e2ff	[B@155dc78	[B@1fd7a27	[B@5eaf0e	[B@36d5dd	[B@12b7df

Figure 3. Encrypted Data



Data Mining Over Encrypted data on Cloud

Data Mining has wide applications in many areas such as banking, medicine, scientific research and among government agencies. Classification is one of the commonly used tasks in data mining applications. With the recent popularity of Cloud computing, users now have the oppurtunity to outsource their data in encrypted form as well as the data mining task to the cloud.

Classification is one of the most commonly used tasks in data mining applications. This technique is used to predict group membership for data instances.

The Customer Details are encrypted and stored in the cloud server. Classification can be applied to these records to recover some useful patterns.

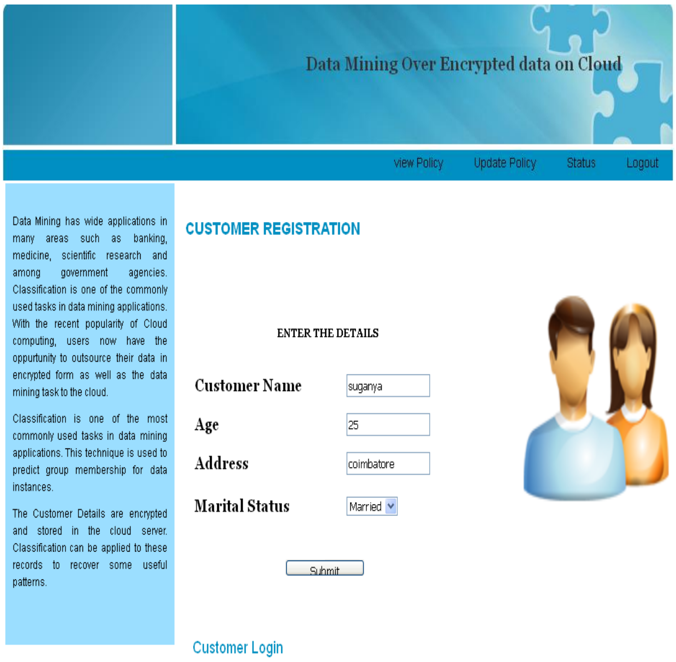
ADD POLICY

POLICY DETAILS

Policy Name	Endowment plans
Type of Level Term	Yearly
Total Amount	10000
Monthly Pay	1000
Profit	5000
Description	maturity benefit. Charge higher fees. Reflected in premiums

Figure 4. Add Policy Details

Fig 2 shows the Admin login page which is initially used for the purpose of Entering in to the page. Fig 4 shows the policy adding details such as amount profit and monthly payout details . Fig 5 shows the customer details and Fig 6 shows the policy view.



Data Mining Over Encrypted data on Cloud

Data Mining has wide applications in many areas such as banking, medicine, scientific research and among government agencies. Classification is one of the commonly used tasks in data mining applications. With the recent popularity of Cloud computing, users now have the opportunity to outsource their data in encrypted form as well as the data mining task to the cloud.

Classification is one of the most commonly used tasks in data mining applications. This technique is used to predict group membership for data instances.

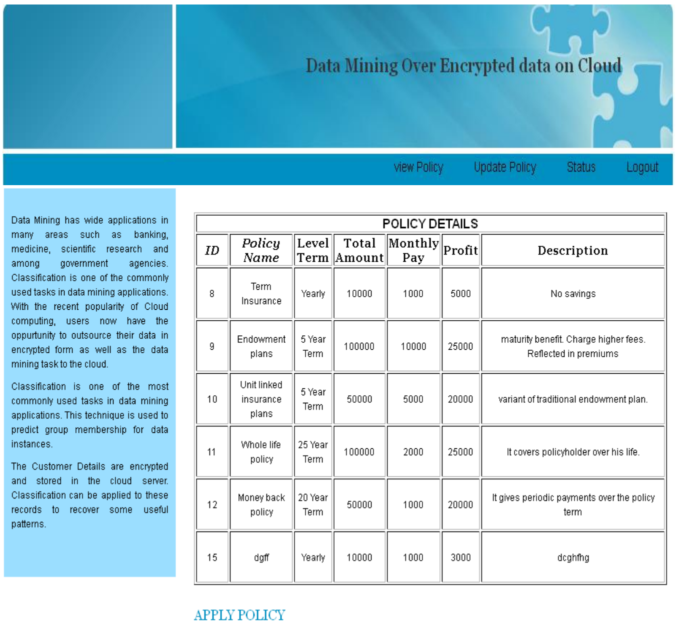
The Customer Details are encrypted and stored in the cloud server. Classification can be applied to these records to recover some useful patterns.

CUSTOMER REGISTRATION

ENTER THE DETAILS

Customer Name	suganya
Age	25
Address	coimbatore
Marital Status	Married

Figure 5. Customer Registration



Data Mining Over Encrypted data on Cloud

Data Mining has wide applications in many areas such as banking, medicine, scientific research and among government agencies. Classification is one of the commonly used tasks in data mining applications. With the recent popularity of Cloud computing, users now have the opportunity to outsource their data in encrypted form as well as the data mining task to the cloud.

Classification is one of the most commonly used tasks in data mining applications. This technique is used to predict group membership for data instances.

The Customer Details are encrypted and stored in the cloud server. Classification can be applied to these records to recover some useful patterns.

POLICY DETAILS

ID	*Policy Name*	Level Term	Total Amount	Monthly Pay	Profit	Description
8	Term Insurance	Yearly	10000	1000	5000	No savings
9	Endowment plans	5 Year Term	100000	10000	25000	maturity benefit. Charge higher fees. Reflected in premiums
10	Unit linked insurance plans	5 Year Term	50000	5000	20000	variant of traditional endowment plan.
11	Whole life policy	25 Year Term	100000	2000	25000	It covers policyholder over his life.
12	Money back policy	20 Year Term	50000	1000	20000	It gives periodic payments over the policy term
15	dgff	Yearly	10000	1000	3000	dcghfhg

Figure 6. View Policy

IV. CONCLUSION

The k-nearest neighbor is one of the commonly used query in many data mining applications. Under an outsourced database environment, where encrypted data are stored in the cloud, secure query processing over encrypted data becomes challenging. In the proposed system, it protects the confidentiality of the data, user's input query, and also hides the data access patterns. The proposed system is mainly implemented to protect the data's in an insurance company. The customer details are encrypted and stored in the cloud server. Only the authenticated user can view the records from the cloud. K-nearest neighbor classification method is used in order to retrieve the records from the cloud server. This classification algorithm can be used to discover useful patterns. Hence the proposed system is more efficient.

V. REFERENCES

[1] P. Mell and T. Grance, "The NIST definition of cloud computing (draft)," NIST Special Publication, vol. 800, p. 145, 2011.

[2] S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "Managing and accessing data in the cloud: Privacy risks and approaches," in Proc. 7th Int. Conf. Risk Security Internet Syst., 2012, pp. 1–9.

[3] P. Williams, R. Sion, and B. Carbunar, “Building castles out of mud: Practical access pattern privacy and correctness on untrusted storage,” in Proc. 15th ACM Conf. Comput. Commun. Security, 2008, pp. 139–148.

[4] P. Paillier, “Public key cryptosystems based on composite degree residuosity classes,” in Proc. 17th Int. Conf. Theory Appl. Cryptographic Techn., 1999, pp. 223–238.

[5] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, “k-nearest neighbor classification over semantically secure encrypted relational data,” eprint arXiv:1403.5001, 2014.

[6] C. Gentry, “Fully homomorphic encryption using ideal lattices,” in Proc. 41st Annu. ACM Sympos. Theory Comput., 2009, pp. 169–178.

[7] C. Gentry and S. Halevi, “Implementing gentry’s fully-homomorphic encryption scheme,” in Proc. 30th Annu. Int. Conf. Theory Appl. Cryptographic Techn.: Adv. Cryptol., 2011, pp. 129 148.

[8] A. Shamir, “How to share a secret,” Commun. ACM, vol. 22, pp. 612–613, 1979.

[9] D. Bogdanov, S. Laur, and J. Willemson, “Sharemind: A framework for fast privacy preserving computations,” in Proc. 13th Eur. Symp. Res. Comput. Security: Comput. Security, 2008, pp. 192–206.

[10] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” ACM Sigmod Rec., vol. 29, pp. 439–450, 2000.

[11] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” in Proc. 20th Annu. Int. Cryptol. Conf. Adv. Cryptol., 2000, pp. 36–54.

[12] P. Zhang, Y. Tong, S. Tang, and D. Yang, “Privacy preserving Naive Bayes classification,” in Proc. 1st Int. Conf. Adv. Data Mining Appl., 2005, pp. 744–752.

[13] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” Inf. Syst., vol. 29, no. 4, pp. 343–364, 2004.

[14] R. J. Bayardo and R. Agrawal, “Data privacy through optimal kanonymization,” in Proc. IEEE 21st Int. Conf. Data Eng., 2005, pp. 217–228.