Implementation of Secure Data Mining Approach in Cloud using Image Encryption

Vijay Kumar
Department of computer
science, Punjab Technical
University CTIEMT,
Jalandhar, Punjab, India

Manju Bala
Department of computer
science, Punjab Technical
University CTIEMT,
Jalandhar, Punjab, India

Pooja Choudhary
Department of computer
science, Punjab Technical
University CTIEMT,
Jalandhar, Punjab, India

ABSTRACT
The continuous development of cloud computing is giving way to more cloud services, due to which security of cloud services, especially data privacy protection, becomes more critical. This research work explores the basic features of data mining techniques in cloud computing and securing the data. The status of the development of cloud computing security, the data privacy analysis, security auditing, data monitoring and other challenges that the cloud computing security faces have been explored. The recent researches on data protection regarding security and privacy issues in cloud computing have partially addressed some issues. The implementation of data mining techniques through cloud computing encourages the users to extract meaningful hidden predictive information from virtually integrated data warehouse that reduces the costs of storage and infrastructure.

Keywords
Cloud Computing, data security, Knowledge Discovery

1. INTRODUCTION
The constant development of information technology in different fields of human life has provoked the broad volumes of information storage in various formats like records, documents, images, sound recordings, videos, scientific data, and many new data formats. The information assembled from diverse applications requires legitimate knowledge/information extraction system to contribute in better decision making. Knowledge discovery in databases (KDD) goes for the exposure of important data from huge accumulations of information. Data mining incorporates numerous methods and algorithms to discover and extract patterns of stored data. From the most recent two decades data mining and knowledge discovery applications have got much attention due to its significance in decision making and it has turn into a vital segment in different associations [1].

Association rule mining is an important research topic of data mining; its task is to find all subsets of items which frequently occur, and the relationship between them. Association rule mining has two main steps: the establishment of frequent itemsets and the establishment of association rules [11].

2. DATA MINING
In this technological era, the data is being generated at an enormous rate. As the advancements in electronics and computer technologies have empowered almost unlimited storage resources, virtually every bit of new data is stored, preserved, and made available. The Internet hosts an almost unimaginable amount of human-generated data across the globe. The endeavours in this direction have brought about various elite data mining libraries, online assets for data search and investigation, cloud services for data storage and analysis, open standards for data exchange and descriptions of data analysis models which portrays the semantics of data, devices and administrations with the mean to empower their interoperability.

The term data mining signifies the action of extracting new, valuable and nontrivial information from extensive volumes of data. The aim is to discover patterns or fabricate models using particular algorithms from various scientific disciplines including artificial intelligence, machine learning, database systems and statistics. The data mining tasks can be classified into two categories with respect to this definition:

2.1 Predictive Data Mining
In predictive data mining, some variables or fields in the database are used to predict unknown or future values of other variables of interest and the goal is to build an executable model from data which can be used for classification, prediction or estimation..

2.2 Descriptive Data Mining
In descriptive data mining, the focus is on finding human-interpretable patterns describing the data and relationships in data.

The KDD process includes an iterative sequence method [3], [4]:

- Selection: The principal step begins with gathering the vital and important information about the domain and setting the objectives to be accomplished.

- Cleaning and Pre-processing: It incorporates discovering mistaken or missing information. It additionally incorporates removal of noise or exceptions along with gathering important data to model or record for noise, representing time arrangement data and known changes.

- Transformation: It is changing over the data into a typical configuration for processing. Data reduction, dimensionality reduction & data transformation method may be used to reduce the number of possible data values being considered.

- Data Mining: This is the most elaborate step as it consists of choosing the function of data mining, choosing the right data mining algorithm and its

application. Choosing the function includes deciding the purpose of the resulting data mining model, such as classification, regression, clustering and summarization.

- Interpretation/Evaluation: In the last step the discovered patterns are evaluated and their validity and relevance are assessed. Redundant and irrelevant patterns are removed while the remaining, relevant patterns are studied and interpreted. Application of the discovered knowledge includes resolving potential conflicts with existing knowledge, taking actions based on the obtained knowledge, such as aiding, modifying and improving existing processes and procedures, especially those involving human experts, and storing, documenting and reporting to interested parties.

3. CLOUD COMPUTING

Cloud Computing is a booming era which guarantees reliable, scalable, pay-per-use, customized and dynamic computing environments for end-users. This new paradigm of cloud computing is appealing vendors and various associations have begun understanding the profits by putting their applications and data into the cloud. This helps in cheaper and efficient utilization of available resources and easier handling of larger computational problems.

Cloud service delivery is divided among three service models-Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS).

The main goal of cloud computing is to combine the distributed resources to achieve higher throughput, high resource utilization and be able to solve large scale computation problems. The cloud computing has many potential advantages in comparison to traditional IT model. But the major barrier for the adoption of cloud computing are the security concerns. Security control measures in cloud are similar to ones in traditional IT environment.

4. SECURITY OF DATA IN CLOUD

Security is a key barrier to the broader adoption of cloud computing. Although cloud computing promises lower costs, rapid scaling, easier maintenance, and service availability anywhere, anytime, a key challenge is how to ensure and build confidence that the cloud can handle user data securely. To make the cloud computing be adopted by users and enterprise, the security concerns of users should be rectified first to make cloud environment trustworthy.

The development of new services bring along new opportunities and challenges. When the data store on personal devices, users have the highest privilege to operate on them and ensure its security. But once the users choose to put data into cloud, they lose their control over the data [9].The user's authentication and authorization is needed to access the data so as to prevent stealing other user's data through service failure or intrusion.

The data stored in the cloud storages is similar with the ones stored in other places and needs to consider three aspects of information security: confidentiality, integrity and availability. The common solution for data confidentiality is data encryption. To ensure the effect of encryption, the use of both encryption algorithm and key strength are needed to be

considered. As the cloud computing environment encompasses large amounts of data transmission, storage and handling so there also needs to consider processing speed and computational efficiency of encrypting large amounts of data. In such cases, symmetric encryption algorithm is more suitable than asymmetric encryption algorithm. The major issue about data encryption is key management. The major issue considered in key management is as who will be responsible for key management. Ideally, the data owners are responsible for managing the key. As the cloud providers need to maintain keys for a large number of users, key management become more complex and difficult [6].

5. STEGANOGRAPHY

The scope of the work is to extract the useful information from large amount of data and store at cloud in secure fashion and then make inferences required by the organization. But the predictions that are generated as a result of mining should be secure from any kind of interception. In this sense, steganography is the best option for sending information secretly because it hides the existence of secret message and provides more security. The security module which is used is image steganography as images are the most popular because of their frequency on the Internet. So the prime focus is to increase the capacity to provide better security during transmission.

Steganography is the process of hiding the one information into other sources of information like text, image or audio file, so that it is not visible to the natural view. In steganography the message is kept secret without any changes but in cryptography the original content of the message is differed in different stages like encryption and decryption. Steganography supports different types of digital formats that are used for hiding the data. These files are known as carriers.

6. APRIORI ALGORITHM

Since there are usually a large number of distinct single items in a typical transaction database, and their combinations may form a very huge number of item sets, it is challenging to develop scalable methods for mining frequent item sets in a large transaction database. The Apriori algorithm is the most general and widely used association rule mining algorithm designed to operate on databases containing transactions. Apriori uses breadth-first-search and a tree structure to count candidate item sets efficiently. It uses an iterative method called layer search to generate (k+1) item sets from k item sets. A k-item set is frequent only if all of its sub-item sets are frequent. This process iterates until no more frequent k-item sets can be generated for some k. This is the essence of the Apriori algorithm [14]-[15].

Apriori Algorithm is based on rule parameters – support, confidence and number of cycles used, but these rule measures are not considered for Predictive Apriori Algorithm. The default number of best rules in Apriori Algorithm and in Predictive Apriori Algorithm is respectively 10 and 100. In Apriori Algorithm, the number of best rules generated is independent of the number of instances and attributes but are dependent on the value minimum support taken.

In Predictive Apriori Algorithm, the best rules depend on the dataset being used and the number of selected attributes. Greater the number of best rules, greater the expected accuracy. A rule is added if the expected predictive accuracy of the particular rule is among 'n' number of best rules and it is not a part of another rule with at least the same expected

predictive accuracy.

7. PROBLEM FORMULATION

The major concerns of users or companies, which put their information on the cloud is they are having no idea what's happening to it. When they will have audit of when their information is approached, who access the data increase to strengthen the confidence that their information is being handled properly. Cloud repository purposes an on-demand information service model, and its reputation increasing because of its scaling down and less repair capital properties. Even, safety measure involvement arises when information repository is overcome to third-party cloud companies. This is essential to able cloud users to check their integrity of the important information on cloud, if the information has corrupted or attacked [9]. Cloud infrastructure is multi-holder, with various applications which are sharing physical framework. That gives aid of much capable resource using. Even there is no physical barriers between them, it is necessary to create and maintain balance safety measure controls to lesser the effect of malwares to distribute via cloud [14].Companies taking cloud services need to understand the involvement for maintaining the confidentiality of owners or other critical business information. The major attention is how the physical location of information affects its use. Ensure only specific users and devices can see sensitive information. One of the biggest concerns for companies coming to contact with cloud computing is confidentiality. In fully-managed public cloud service, confidentiality and aloofness risks are often likely to change accordingly to the provider's aloofness policy.

8. PROPOSED SCHEME

The paper aims to:

1. Implement secure cloud system using CloudSim simulator and java Eclipse

2. Collection and preprocessing of data for mining

3. Encrypt the dataset into an image and securely migrate that image to cloud

4. Apply data mining in cloud and secure the generated mining report.

The proposed methodology is as:
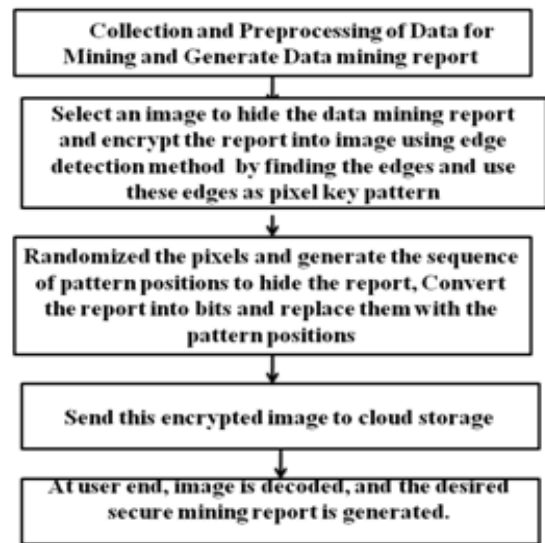
9. RESULTS AND DISCUSSION

This section presents the simulation results of the proposed system implemented in CloudSim simulator and Java Eclipse.
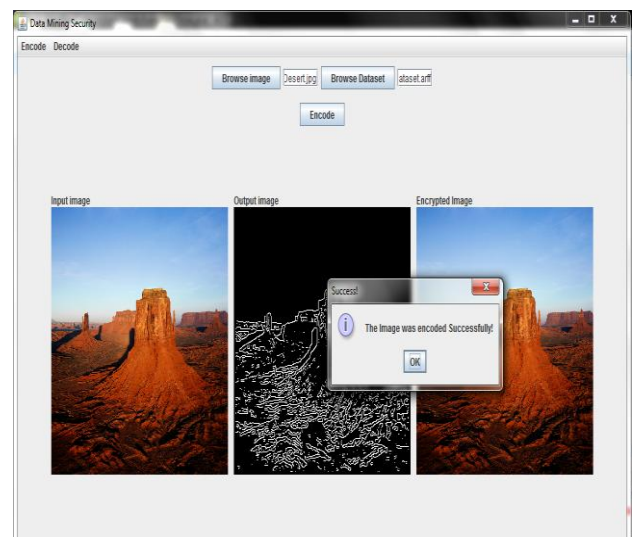

Fig.1 Image Encoded

```
Msg is
Apriori
=======

Minimum support: 0.7 (1196 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:

1. LastInterPurchaseTimeDifference=high 1346 ==> ProductCategory=norm 1323    conf:(0.98)
2. RelativeMonetaryValue=high LastInterPurchaseTimeDifference=high 1242 ==> ProductCategory=norm 1219    conf:(0.98)
3. ProductCategory=norm 1375 ==> LastInterPurchaseTimeDifference=high 1323    conf:(0.96)
4. RelativeMonetaryValue=high ProductCategory=norm 1271 ==> LastInterPurchaseTimeDifference=high 1219    conf:(0.96)
5. RelativeMonetaryValue=high 1340 ==> ProductCategory=norm 1271    conf:(0.95)
6. RelativeMonetaryValue=high 1340 ==> LastInterPurchaseTimeDifference=high 1242    conf:(0.93)
7. ProductCategory=norm 1375 ==> RelativeMonetaryValue=high 1271    conf:(0.92)

Process  complete
TIME To Decode the result: 60632 ms
```

Fig.2 Decoded message from cloud and decode time

```
APRIORI RESULT:

Apriori
=======

Minimum support: 0.7 (1196 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 6

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:

1. LastInterPurchaseTimeDifference=high 1346 ==> ProductCategory=norm 1323    conf:(0.98)
2. RelativeMonetaryValue=high LastInterPurchaseTimeDifference=high 1242 ==> ProductCategory=norm 1219    conf:(0.98)
3. ProductCategory=norm 1375 ==> LastInterPurchaseTimeDifference=high 1323    conf:(0.96)
4. RelativeMonetaryValue=high ProductCategory=norm 1271 ==> LastInterPurchaseTimeDifference=high 1219    conf:(0.96)
5. RelativeMonetaryValue=high 1340 ==> ProductCategory=norm 1271    conf:(0.95)
6. RelativeMonetaryValue=high 1340 ==> LastInterPurchaseTimeDifference=high 1242    conf:(0.93)
7. ProductCategory=norm 1375 ==> RelativeMonetaryValue=high 1271    conf:(0.92)

Time taken to generate the rules by APRIORI : 132 ms
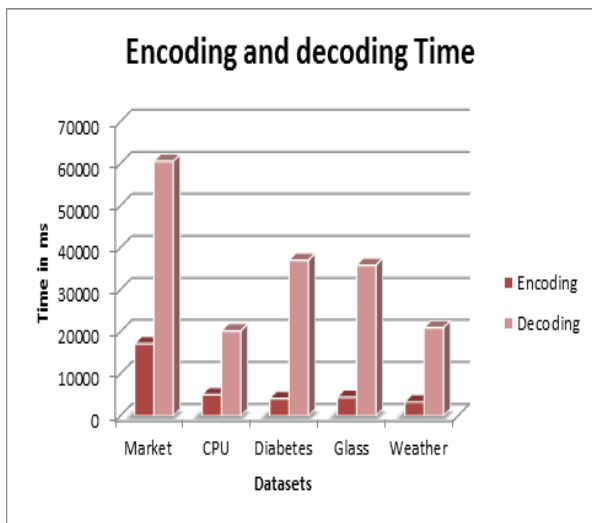```

**Fig.3 Apriori Results**



**Fig.4 Encoding/ Decoding Time Analysis**

## 10. CONCLUSION

In an emerging discipline like cloud computing, security needs to be analyzed more frequently. With advancement in cloud technologies and increasing number of cloud users, data security dimensions will continuously increase. Cloud computing security needs consider both technology and strategy, including: audit, compliance and risk assessment. Both the Service providers and the clients must work together to ensure safety and security of cloud and data on clouds. Mutual understanding between service providers and users is extremely necessary for providing better cloud security. In our paper we are laying stress on the security issue in the cloud. The paper presents the simulation results.

## 11. REFERENCES

[1] UppunuthulaVenkateshwarlu, PuppalaPriyanka, "Survey on Secure Data mining in Cloud     Computing", ISSN : 2347 - 8446 (Online) ISSN : 2347 - 9817 (Print), International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2, Issue 2, Ver. 1 (April - June 2014)

[2] Juan Li1, Pallavi Roy, Samee U. Khan, Lizhe Wang, Yan Bai, "Data Mining Using Clouds: An Experimental Implementation of Apriori over MapReduce", The 12th IEEE International Conference on Scalable Computing and Communication, December, 2012.

[3] Pramod Kumar Joshi1 and Sadhana Rana, "Era of Cloud Computing", High Performance Architecture and Grid Computing Communications in Computer and Information Science, Vol. 169,, pp 1-8, ISSN 1865-0929, Springer-Verlag Berlin Heidelberg 2011

[4] EmanElghoniemy, OthmaneBouhali, Hussein Alnuweiri, "Resource Allocation and Scheduling in loud Computing", DOI: 978-1-4673-0009-4/12, IEEE 2012

[5] Rabi Prasad Padhy, ManasRanjanPatra, Suresh Chandra Satapathy, "Cloud Computing: Security Issues and Research Challenges", International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 1, No. 2, December 2011

[6] Prof. G. Thippa Reddy, Prof. K. Sudheer , Prof. K Rajesh, Prof. K. Lakshmanna, "Employing Data Mining On Highly Secured Private Clouds For Implementing A Security-As a- Service Framework", Journal of Theoretical and Applied Information Technology, Vol. 59 No.2, ISSN: 1992-8645, January20, 2014

[7] Ramadhan Mstafa1, Christian Bach, "Information ,Hiding in Images Using Steganography Techniques", ASEE Northeast Section Conference Norwich University,Reviewed Paper, March 14-16, 2013

[8] Cong Wang, Qian Wang, and KuiRen,Wenjing Lou, "Ensuring Data Storage Security in Cloud Computing", Quality of Service, 2009. IWQoS. 17th International Workshop, ISSN : 1548-615X, DOI: 10.1109/IWQoS.2009.5201385, IEEE, July 13-15, 2009

[9] C.P.Sumathi1, T.Santanam2 and G.Umamaheswari, "A Study of Various Steganographic Techniques Used for Information Hiding", International Journal of Computer Science & Engineering Survey (IJCSES) Vol.4, No.6, December 2013

[10] Monjur Ahmed1 and Mohammad Ashraf Hossain, "Cloud computing and security issues in the cloud" , International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.1, January 2014

[11] Lingjuan Li Min Zhang, " The Strategy of Mining Association Rule Based on Cloud Computing",Business Computing and Global Informatization (BCGIN), International Conference, DOI: 10.1109/BCGIn.2011.125, IEEE, July29-31, 2011

[12] Zhangn Chun-sheng Li yan, "Extension of Local Association Rules Mining Algorithm Based on Apriori Algorithm",  DOI: 978-1-4799-3279-5 /14IEEE, 2014

[13] B. Kamala, "A study on integrated approach of data mining and cloud mining", International Journal of Advances In Computer Science and Cloud Computing, ISSN: 2321-4058 Vol. 1, Issue- 2, Nov-2013

[14] Zeba Qureshi1, Jaya Bansal2, Sanjay Bansal3, "A Survey on Association Rule Mining in Cloud Computing", International Journal of Emerging Technology and Advanced Engineering , ISSN 2250-2459, Vol. 3, Issue 4, April 2013

[15] Jiawei Han, Hong Cheng, Dong Xin, "Xifeng Yan Frequent pattern mining: current status and future", , Data Min Knowl Disc 15:55–86, DOI 10.1007/s10618-006-0059-1, 2007

[16] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, 0738-4602-1996

[17] VahidAshktorab, Seyed Reza Taghizadeh , "Security Threats and Countermeasures in Cloud computing", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol. 1, Issue 2, ISSN 2319 – 4847, October 2012

[18] GarimaSaini Naveen Sharma, "Triple Security of Data in Cloud Computing", International Journal of Scientific and Research Publications, Vol. 4, Issue 6, ISSN 2250-3153, June 2014

[19] Jijo.S. Nair, BholaNath Roy, " Data Security in Cloud", International Journal of Computational Engineering Research (IJCER) ISSN: 2250-3005, National Conference on Architecture, Software system and Green computing

[20] T.V. Mahendra , N.Deepika , N.KeasavaRao, "Data Mining for High Performance Data Cloud using Association Rule Mining", International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 2, Issue 1, ISSN: 2277 128X,  January 2012

[21] Sneha Arora1, Sanyam Anand2, "A New Approach for Image Steganography using Edge Detection Method", International Journal of Innovative Research in Computer and Communication Engineering  Vol. 1, Issue 3,  ISSN (Print) : 2320 – 9798 ISSN (Online): 2320 – 9801, May 2013