International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4, 2025
https://theaspd.com/index.php

Designing and Developing Utility Based Privacy Preserving System Using Data Mining Over Big Data in Cloud Systems and Environments.

Dr.G.Siva Nageswara Rao[1], Mekala Bhanu Venkata Yeswanth Reddy[2],
[1]Professor**,** Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India**.**
sivanags@kluniversity.in.
[2]PG student **,** Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India**.**
yeswanth1908@gmail.com.

Abstract
The rapid growth of big data and its integration into cloud systems has revolutionized data- driven decision-making. However, the inherent privacy risks associated with data storage, sharing, and processing have become a significant challenge. This study focuses on designing and developing a utility-based privacy-preserving system that leverages data mining techniques to safeguard sensitive information while maintaining data usability in cloud environments. The proposed system employs advanced anonymization, encryption, and differential privacy mechanisms to ensure data confidentiality without compromising analytical accuracy. By utilizing scalable data mining algorithms, the system is capable of handling the complexities and volume of big data in a distributed cloud infrastructure. Furthermore, the system incorporates a utility-driven model to balance privacy preservation with data utility, enabling organizations to extract meaningful insights while adhering to regulatory and ethical standards. Experimental evaluations on real-world datasets demonstrate the system's effectiveness in mitigating privacy risks while retaining high levels of data utility. This innovative approach provides a robust framework for privacy-preserving data mining in cloud systems, addressing the growing concerns of data security and privacy in the era of big data. The proposed solution has practical applications in domains such as healthcare, finance, and e-governance, where privacy and utility are critical.

INTRODUCTION AND PROBLEM IDENTIFICATION:

Introduction to DM, Big Data and Cloud Computing:

In recent years, data mining (DM) has attracted more and more attention, probably because of the popularity of the "big data" concept. On other hand, cloud computing provides massive computation power and storage capacity which enable users to deploy applications without infrastructure investment. Coupled with cloud computing, data sets have become so large and complex that it is a considerable challenge for traditional data processing tools to handle the analysis pipeline of these data. Normally, such data sets are often from various sources and of different types (Variety) such as unstructured social media content and half-structured medical records and business transactions, and are of large size (Volume) with fast data in/out (Velocity). The Map Reduce framework has been widely adopted by a large number of companies and organizations to process huge-volume data sets [10] (2010). A typical example is the Amazon Elastic Map Reduce (Amazon EMR) service [26] (2013). In this way, it is economical and convenient for companies and organizations to capture, store, organize, share and analyze Big Data to gain competitive advantages. However, privacy concerns in Map Reduce platforms are aggravated because the privacy-sensitive information scattered among various data sets can be recovered with more ease when data and computational power are considerably abundant. Although some privacy issues are not new, their importance is amplified by cloud computing and Big Data [12] (2012). Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services [1](2013). Cloud services include the delivery of software, infrastructure, and storage over the internet either as separate components or a complete platform based on user demand [2] (2014). Cloud computing does not require a user to be in a specific place to gain access to it. Companies may find that cloud computing allows them to reduce the cost of information management, since they are not required to own their own servers and can use capacity leased from third parties. Moreover, Cloud computing needs to address three main security issues: confidentiality, integrity

and availability. Due to the greater level of flexibility, the cloud has become the proliferating ground of a new generation of products and services.

However, the flexibility of services of cloud imposes the risk of the security and privacy of users' data [3](2009). Cloud applications such as data storage, data retrieval and data portability have become some significant needs for IT organizations dealing with cloud computing. Considering the requirement, the IT development and user oriented global services can be globalized and delivered to single click by means of cloud applications such as Big Data [4] (2014). Therefore, many companies or organizations have been migrating or building their business into cloud.

Problem Identification:

However, numerous potential customers are still hesitant to take advantage of cloud due to security and privacy concerns [5] (2010), [6] (2011) Nevertheless, providing such secure and privacy-preserving data services is very challenging, as security problems can arise in multiple levels of the data services, and security and privacy protection may impede functionality and performance of the data services [7] (2012). Moreover, Privacy-preserving techniques like generalization can withstand most privacy attacks on one single dataset, while preserving privacy for multiple datasets is still a challenging problem.Thus, for preserving privacy of multiple datasets, it is promising to anonymize all datasets first and then encrypt them before storing or sharing them in cloud. Usually, the volume of intermediate datasets is huge [8] (2013). When the amount of data is large, it is difficult or even impossible for the data to be stored on a single machine, which renders sequential algorithms unusable. In situations where the amount of data is prohibitively large, the Map Reduce [9] (2004) programming paradigm is used to overcome this obstacle. Fortunately, the Map Reduce framework proposed by Google [10] (2010), which simplifies the programming for distributed data processing, and the Hadoop implementation by Apache Foundation [11] (2013), which makes the framework freely available for everyone, distributed programming is becoming more and more popular.

However, privacy concerns in Map Reduce platforms are aggravated because the privacy-sensitive information scattered among various data sets can be recovered with more ease when data and computational power are considerably abundant.

Although some privacy issues are not new, their importance is amplified by cloud computing and Big Data [12] (2012).

Recently, the research on privacy issues in the Map Reduce framework on cloud has commenced. Mechanisms such as:

encryption [13] (2011),

access control [14] (2011),

differential privacy [15] (2010) and

auditing [16] (2011) are exploited to protect the data privacy in the Map Reduce framework.

These mechanisms are well-know pillars of privacy protection and still have open questions in the context of cloud computing and Big Data.

If we encrypt these data sets, processing on data sets efficiently will be quite a challenging task, because most existing applications only run on unencrypted data sets. Furthermore, lot of encrypted algorithms also developed to big data privacy. Data anonymization is a promising category of approaches to achieve such a goal

[17] (2010). However, the computing infrastructure and paradigm has been moving to the Map Reduce framework in order to get scalability, e.g., the newly emerging project Apache Mahout [18] (2013).

Thus, how to achieve privacy preservation and high utility of Big Data in the Map Reduce framework on cloud for mining or analytic applications is still a challenge problem and needs extensive investigation.

Despite that the information discovered by data mining can be very valuable to many applications; people have shown increasing concern about the other side of the coin, namely the privacy threats posed by data mining.

Individual's privacy may be violated due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc.

To deal with the privacy issues in data mining, a subfield of data mining, referred to as privacy preserving data mining (PPDM) has gained a great development in recent years.
The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The process of PPDM leads to loss of information for data mining purposes.
This loss of information can also be considered a loss of utility for data mining purposes. Since some negative results [25] (2005) on the curse of dimensionality
suggest that a lot of attributes may need to be suppressed in order to preserve anonymity, it is extremely important to do this carefully in order to preserve utility.

LITERATURE SURVEY:

In, Yogachandran *et al.* [19] (2014) have explained the Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud. privacy- preserving (PP) data classification technique where the server was unable to learn any knowledge about clients' input data samples while the server side classifier was also kept secret from the clients during the classification process. More specifically, they explained the first known client-server data classification protocol using support vector machine. The protocol performs PP classification for both two-class and multi-class problems. The protocol exploits properties of Pailler homomorphic encryption and secure two-party computation. At the core of the protocol lies an efficient, novel protocol for securely obtaining the sign of Pailler encrypted numbers. Moreover, Xuyun Zhang *et al.* [20] (2013) have explained the Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud. Encrypting ALL data sets in cloud was widely adopted in existing approaches to address this challenge. But they argue that encrypting all intermediate data sets were neither efficient nor cost-effective because it was very time consuming and costly for data-intensive applications to en/decrypt data sets frequently while performing any operation on them. Here, they explained a upper bound privacy leakage constraint-based approach to identify which intermediate data sets need to be encrypted and which do not, so that privacy-preserving cost was saved while the privacy requirements of data holders was satisfied. Evaluation results demonstrate that the privacy-preserving cost of intermediate data sets was significantly reduced with our approach over existing ones where all data sets were encrypted. Likewise, Xingjian Li [21] (2014) have explained the Mining Frequent Item sets from Library Big Data. Frequent itemset mining plays an important part in college library data analysis. Because there was a lot of redundant data in library database, the mining process may generate intra-property frequent itemsets, and this hinders its efficiency significantly. To address this issue, they introduced an improved FP-Growth algorithm they call RFP-Growth to avoid generating intra-property frequent itemsets, and to further boost its efficiency, implement its Map Reduce version with additional prune strategy. The algorithm was tested using both synthetic and real world library data, and the experimental results showed that the algorithm outperformed existing algorithms.In, Chandramohan Dhasarathan *et al.* [22] (2015) have explained the preserve user information privacy for a pervasive and ubiquitous environment. Preserving proprietor's data and information in cloud was one of the top most challenging missions for cloud provider. Here they explained a hybrid authentication technique as an end point lock. It was a composite model coupled with an algorithm for user's privacy preserving, which was likely to be Hash Diff Anomaly Detection and Prevention (HDAD). This algorithmic protocol acts intelligently as a privacy preserving model and technique to ensure the user's data were kept more secretly and develop an endorsed trust on providers. They also explore the highest necessity to maintain the confidentiality of cloud user's data.Moreover, Xin Dong *et al.* [23] (2014) have explained the privacy-preserving data sharing service in cloud computing. They explained an effective, scalable and flexible privacy-preserving data policy with semantic security, by utilizing ciphertext policy attribute-based encryption (CP-ABE) combined with identity-based encryption (IBE) techniques. In addition to ensuring robust data sharing security, their policy succeeds in preserving the privacy of cloud users and supports efficient and secure dynamic operations including, but not limited to, file creation, user revocation and modification of user attributes. Security analysis indicates that the system policy was secure under the generic bilinear group model in the random oracle model and enforces

fine-grained access control, full collusion resistance and backward secrecy. Furthermore, performance analysis and experimental results was show that the overhead was as light as possible.Additionally, Xuyun Zhang *et al.* [24] (2012) have explained the Privacy-Preserving Layer over Map Reduce on Cloud. Cloud computing provides powerful and economical infrastructural resources for cloud users to handle ever increasing Big Data with data- processing frameworks such as Map Reduce. Based on cloud computing, the Map Reduce framework was widely adopted to process huge-volume data sets by various companies and organizations due to its salient features. Nevertheless, privacy concerns in Map Reduce were aggravated because the privacy-sensitive information scattered among various data sets was recovered with more ease when data and computational power were considerably abundant. Existing approaches employ techniques like access control or encryption to protect privacy in data processed by Map Reduce. However, such techniques fail to preserve data privacy cost-effectively in some common scenarios where data were processed for data analytics, mining and sharing on cloud. As such, they explained a flexible, scalable, dynamical and cost effective privacy-preserving layer over the Map Reduce framework in this paper. The layer ensures data privacy preservation and data utility under the given privacy requirements before data were further processed by subsequent Map Reduce tasks. A corresponding prototype system was developed for the privacy-preserving layer as well.

Notable Generic Issues Identified:

Data sets have become so large and complex, it is a considerable challenge for traditional data processing tools to handle the analysis pipeline of these data.Individual's privacy may be violated due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc.

Specific Issues Identified:

Considering above said issues, an efficient privacy preserving data mining (PPDM) technique is urgently needed and utmost important and so much in demand.

There is no utility based privacy preserving system, data mining over Big data in cloud systems and also in cloud environments.

It is observed from survey efficient optimization techniques are not available to deal with multiple datasets.

In the literature survey, several methods have been proposed for the Privacy Preserving of big data in Cloud, but the complete privacy preserving system is missing, creating the same is utmost important and so much in demand.

PROPOSED RESEARCH OBJECTIVES:

Detailed literature survey will be conducted on Data Mining, Big data and Cloud Computing Fields to understand the strengths and weaknesses and also to enrich the proposed work.

Study of different data mining techniques, Big data analytics and Cloud Computing methods and techniques available and also utilized for Privacy Preserving.

To design effective system or model using latest techniques and methods for Privacy Preserving-Aware on Big Data in Cloud environments, To conduct experimental validation and analysis of the designed utility based Privacy Preserving based data mining (PPDM) system.

To create effective Optimization for preserving data utility framework.

Designing and developing utility based privacy preserving system using data mining over big data in cloud systems and environments.

PROPOSED METHODOLOGY AND TECHNIQUES:

The main intention of this research is to develop a utility based privacy preserving data mining (PPDM) over big data in cloud systems.

The utility-based privacy preservation has two goals:

protecting the private information

preserving the data utility as much as possible.

Through the literature survey, we understand that utility based privacy preserving data mining (PPDM)

goals are to be reformulated.

Differential Privacy

Differential Privacy ensures that the output of a computation is indistinguishable whether a particular individual's data is included or not:

$Pr[M(D1) \in S] \leq e^{\varepsilon} \cdot Pr[M(D2) \in S]$

Where:

M: Privacy-preserving mechanism.

D1, D2: Databases differing by one record.

S: Subset of possible outputs.

ε: Privacy budget controlling the privacy-utility trade-off. Data Utility Metric

Utility is measured as the difference in accuracy between the original and transformed datasets:

Utility = 1 - (Error_transformed / Error_original) k-Anonymity

A dataset satisfies k-anonymity if each record is indistinguishable from at least k-1 other records based on quasi-identifiers:

$\forall t \in T, freq(t) \geq k$

Where:

T: Set of records.

freq(t): Frequency of record t. l-Diversity

l-Diversity ensures sensitive attributes have at least l 'well-represented' values in each equivalence class:

$\forall EC \in T, H(EC) \geq \log_2(l)$

Where:

EC: Equivalence class.

H(EC): Entropy of sensitive attributes in EC.

Big Data Distribution in Cloud Systems

For scalability, the total processing time Tp in a distributed cloud system can be represented as: $Tp = \max\{i=1 \text{ to } n\} (Tc_i + Tn_i)$

Where:

n: Number of nodes.

Tc_i: Computation time on node i.

Tn_i: Network transfer time on node i.

Privacy-Utility Trade-off

The privacy-utility trade-off can be modeled as: $U = f(\varepsilon)$ subject to $\varepsilon \leq \varepsilon_{max}$

Where:

U: Data utility.

f(ε): Utility function based on privacy budget ε.

ε_max: Maximum allowable privacy budget.

Techniques:

A convolution process will be done in our research between a new kernel matrix and big data. The new kernel matrix will be designed and optimally find out through hybrid algorithm (GABC). In GABC, genetic and artificial bee colony algorithm will be combined and used for hybridization process.

Two phases of privacy-persevering framework over big data in cloud systems will be performed.

In the first phase, an efficient classifier based utility measure will be developed using radial basis function-neural network (RBF-NN), which should capture the intrinsic factors that affect the quality of data for our application.

In the second phase, A Map reduce framework will be proposed to protect the private information, which is responsible for anonymizing original data sets according to privacy requirements.

The system will be implemented using Java, R , Python and evaluated under standard metrics

POSSIBLE OUTCOMES:

Efficient utility based privacy preserving data mining (PPDM) over big data in cloud systems can be achieved by our implementation results and discussions

Privacy information will be well secured through our proposed system.
Detailed literature survey report through this research work will be valuable pathway to the future research work in the field of Privacy Preserving.
Different data mining techniques, Big data analytics and Cloud Computing methods and techniques identified and utilized for Privacy Preserving in our proposed work can be utilized for further and future experiments in all related areas.
Design which we are going to suggest through our work will create effective system or model using latest techniques and methods for Privacy Preserving-Aware on Big Data in Cloud environments
Experimental validation and analysis of the designed utility based Privacy Preserving based data mining (PPDM) system .
Our research work will provide Optimization for preserving data utility framework.
Designing and developing utility based privacy preserving system using data mining over big data in cloud systems and environments will be achieved.

REFERENCES:

1. Jaydip Sen, "Security and Privacy Issues in Cloud Computing", 2013.
2. Kaustubh Satpute,Charudatt Satpute and Dipti Bhade, "Review on Internet base Services of Cloud Computing", International Journal on Recent and Innovation Trends in Computing and Communication ,vol. 2 no. 2 ,2014.
3. Introduction to Cloud Computing Architecture by Sun Microsystems, Inc., june 2009.
4. Chhaya S Dule, H.A. Girijamma and K.M Rajasekharaiah, "Privacy Preservation Enriched MapReduce for Hadoop Based BigData Applications",American International Journal of Research in Science, Technology, Engineering & Mathematics, vol.6, no.3, pp. 293-299, 2014
5. H. Takabi, J.B.D. Joshi, and G. Ahn, "Security and Privacy Challenges in Cloud Computing Environments", IEEE Security & Privacy, vol. 8, no. 6, pp. 24-31, Nov./Dec. 2010.
6. D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues", Future Generation Computer Systems, vol. 28, no. 3, pp. 583- 592, 2011.
7. Divyakant Agrawal, Amr El Abbadi and Shiyuan Wang, "Secure and Privacy-Preserving Data Services in the Cloud: A Data Centric View",Proceedings of the VLDB Endowment, Vol. 5, No. 12, 2012
8. S.Hemalatha and S.Alaudeen Basha, "Enabling for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud", International Journal of Scientific and Research Publications, vol. 3, no. 10, 2013.
9. Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In Proceedings of OSDI, pages 137–150, 2004.
10. Dean J, Ghemawat S (2010) Map Reduce: a flexible data processing tool. Commun ACM 53(1):72–77.
11. Apache hadoop. http://hadoop.apache.org/, 2013.
12. S. Chaudhuri, "What next?:Ahalf-dozen data management research goals for big data and the cloud" In: Proceedings of the 31st symposium on principles of database systems, pp. 1–4, 2012.
13. N.Cao, C. Wang , Li M, Ren K, Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", In: Proceedings of the 31st annual IEEE international conference on, computer communications, pp. 829–837, 2011.
14. Puttaswamy KPN, Kruegel C, Zhao, "toward data confidentiality in storage-intensive cloud applications", in: Proceedings of the 2nd ACM symposium on cloud computing (SoCC'11), article 10
15. Roy I, Setty STV, Kilzer A, Shmatikov V, Witchel , ""security and privacy for MapReduce", In: Proceedings of 7th USENIX conference on networked systems design and implementation, pp 297–312, 2010.
16. Xiao Z, Xiao Y , "Accountable Map Reduce in cloud computing", In Proceedings of the 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS), pp1082–1087.
17. Fung BCM, Wang K, Chen R, Yu PS, "Privacy-preserving data publishing: a survey of recent developments", ACM Computer Survey, vol.42, no.4, pp.1-53, 2010.
18. Apache (2013) Apache Mahout machine learning library. http://mahout.apache.org/.
19. Accessed on 10 Mar 2013
20. Yogachandran Rahulamathavan, Raphael C.-W. Phan,Suresh Veluru, Kanapathippillai Cumanan and Muttukrishnan Rajarajan, "Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud", IEEE transactions on dependable and secure computing, vol. 11, no. 5, 2014.
21. Xuyun Zhang, Chang Liu, Surya Nepal, Suraj Pandey and Jinjun Chen, "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud", IEEE Transactions On Parallel And Distributed Systems, Vol. 24, No. 6, Pp. 1192-1203, 2013.
22. Xingjian Li, "An Algorithm for Mining Frequent Itemsets from Library Big Data", journal of software, vol. 9, no. 9, 2014.
23. Chandramohan Dhasarathan, Sathian Dananjayan, Rajaguru Dayalan, Vengattaraman Thirumal and Dhavachelvan Ponnurangam, "A multi-agent approach: To preserve user
24. information privacy for a pervasive and ubiquitous environment", Egyptian Informatics Journal, vol. 16, pp.151–166, 2015.
25. Xin Dong, Jiadi Yu, Yuan Luo, Yingying Chen, Guangtao Xue and Minglu Li, "Achieving an effective, scalable and privacy-