PRAGMATIC PRIVACY SECURE IN EMPLOYEE MANAGEMENT SYSTEM USING MAP REDUCE BASED K-MEANS CLUSTERING OVER LARGE-SCALE DATASET

C.SELVI *1, J.MALATHI *2, M.MOUNIKA*3, K.RAJAMANI*4

1,3*PG Student, Department of Computer Applications, New Prince Shri Bhavani College of Engineering and Technology, Gowrivakkam, Chennai, Tamilnadu, India.
3,4Assistant Professor, Department of Computer Science, New Prince Shri Bhavani College of Engineering and Technology, Gowrivakkam, Chennai, Tamilnadu, India.
4rajamani@newprinceshribhavani.com

Abstract: Privacy secure has been adopted in many real-world data analysis applications such as, data mining, document retrieval, image annotation and pattern classification. The data is taken from cloud server to encrypted dataset retrieve the database. It is widely used in voluted security and privacy in a service model, and need to protect the data running on the platform. Through the concepts it contingents the privacy secure k-means clustering over large scale dataset using Employee Management System. For an Encrypted dataset, reveals the method as security analysis and numerical analysis.

Keywords: Privacy preserving K-means Clustering, security, Hadoop, Bigdata, cloud computing.

1. INTRODUCTION

Grouping of Data is one of the major task of exploratory data mining and statistical data analysis, it is adopted many domains. Using the K-means clustering algorithm based on MapReduce. The big data is used to describe huge datasets having the four V’s are volume, variety, velocity and value. The problem of privacy secure k-means clustering has been investigated under the multiparty secure computation model, in which owners of distributed datasets interact for clustering without disclosing their own datasets to each other. The k-means clustering algorithm is a method of cluster analysis as partitions n objects into k clusters. The clusters are formed to enhance a detached partitioning criterion based on a correlation function, such as distance and that objects within a cluster are ‘‘similar’’, otherwise the objects of different clusters are ‘‘dissimilar’’ in terms of the database aspects. The privacy-secure k-means clustering
has been mainly achieved by either data perturbation or secured multiparty computation. Input perturbation and output perturbation, are the first input perturbation techniques trail to protect the data sites privacy by having each data site. Second is output perturbation techniques are build up done by statistical database.

MapReduce is a programming model and an associated implementation for processing and generating large scale datasets that is variety of real-world tasks. The users specify the computation in terms of a map and a reduce function, and the runtime system automatically parallelizes the computation across large-scale clusters of machines, and handles machine failures, and schedules inter-machine communication to make efficient use of the network and disks. The google and Hadoop provide to MapReduce runtimes with fault tolerance and dynamic flexibility support. Privacy secure data mining provides a secure way to compute the output of a particular algorithm applied to the union of privately distributed data sets without sharing them. The commonly known data mining algorithms have been modified to preserve the privacy of distributed data sets, such as decision-tree learning, association rule mining, naive Bayes, and support vector machines. Other machine learning algorithms, and reinforcement learning or belief propagation has been modified for privacy preservation, these algorithms assume that public or private organizations that collect a large amount of personal information are Privacy secure data mining participants. They are responsible for privacy secure, and the Privacy preserving data mining is conducted only when these organizations reach an agreement. We designate this framework as server-centric privacy preservation. Server-centric privacy preservation mainly assumes computations among a relatively small number of participants and is well-suited to data mining between enterprises.

2. EXISTING WORK

The clustering is one of the major task of exploratory data mining and statistical data analysis, which has been ubiquitously adopted in many domains, and including healthcare, social network, image analysis. The rapid growth of big data involved in today’s data mining and analysis also introduces challenges for clustering over them in terms of volume, variety, and velocity. It was convenient managing the large-scale datasets and support clustering. The public cloud infrastructure was acting the major role for both performance and economic consideration. The Weighted k-means clustering used by secure on small amount of data.
Limitations:

- It doesn’t provide security.
- Threat to the personal information

3. PROPOSED WORK

The prospective method of practical privacy-secure K-means clustering scheme that are efficiently outsourced to cloud servers. This scheme allows the cloud servers to perform on clustering directly over encrypted datasets, and also achieving comparable computational ramification and accuracy, when it’s compare to clustering over unencrypted data. It also Authenticates Secure Integration of MapReduce into the scheme, that is extremely suitable for cloud computing environment. Through all this security analysis and numerical analysis, it carries the performance of the scheme in terms of security and efficiency. The intensive security examination and numerical Authentication can act as a plan regarding security and productivity. We use some technologies like Honey encryption key, and HQL query is used for storing purpose. The proposed system has more than large amount of data set are grouped into split the data to be encrypted.

4. RELATED WORK

The Advanced appearance of this Employee Management system has to be added some special modules and various characteristics and Techniques. Xun Yi et al [1] describe about the Equally contributory privacy-preserving k-means clustering over vertically partitioned data, in which each party equally contributes to k-means clustering. They were built on ElGamal’s encryption algorithm. Geetha Jagannathan[2], describes about Privacy preserving distributed k-means clustering over arbitrary partitioned data., They can provide two major contributions in privacy-preserving data mining. The first is Introduce the concept of arbitrarily partitioned data, which is a generalization of both horizontally and vertically partitioned data. Second, it’s provide an efficient privacy-preserving protocol for k-means clustering in the setting of arbitrarily partitioned data. It can be realized based on secure homomorphic encryption scheme. The examples of privacy preserving are used by arbitrary partitioned data on Alice and bob Encryption methods.
Paul Bunn et al [3]. They can explain a Two-Party k-Means Clustering Protocol that guarantees privacy, and is more efficient than utilizing a general multiparty “compiler” to achieve the same task, and used two main techniques are secure homomorphic encryption and Paillier Homomorphic Encryption. Weizhong Zhao [4]. describing about fast Parallel k-means clustering based on MapReduce, which has been widely embraced by both academia and industry. They use speedup, scaleup and size up to evaluate the performances of their project algorithm. Wai Kit Wong [5]. Describe about the Secure knn computation on encrypted databases as can be developed securely support kNN applications under the SCONEDB model. An approach is used for securely support distance preserving transformation (DPT). They develop a new asymmetric scalar-product-preserving encryption (ASPE).

Ning Cao [6], has been describe the Privacy-preserving query over encrypted graph-structured data in cloud computing. It’s used for if any negative or duplicate data are store in the database, that’s check before store the data. It has using filtering and verification techniques. Jun Sakuma [7], detailed about Large-scale k-means clustering with user-centric privacy-preservation, the use of this paper is update cluster center and update cluster labels, and the k-means algorithm combined with a decentralized cryptographic protocol and a gossip-based protocol. Like used as methods is decision-tree learning, association rule mining, k-means clustering essentially includes two key steps. Privacy-preserving gossip-based aggregation Private and decentralized determination of the nearest neighbor.

5. DATA ANALYTIC MODULE WITH PIG

The figure 1 shows the Apache Pig which is a high-level data flow platform for execution on Map Reduce programs. Pig handles both structure and unstructured language. It is also top of the map reduce process running background. In this module also used for analyzing the Data set through Pig using Latin Script data flow language, in this also we are doing all operators, functions and joins applying on the data see the result.
6. DATA ANALYTIC MODULE WITH MAPREDUCE

The figure 2 shown in data analytic module with MapReduce is handle some approaches and a program model for shared computing established on java. The MapReduce methods contain two important tasks, as Map and Reduce. In this module also used for analyzing the data set using MAP REDUCE. Map Reduce is run by Java Program.

7. ARCHITECTURE DIAGRAM

The system specifies the two major entities which are shown in the figure.3 which includes the client (dataset owner) and cloud server. The dataset owner has collection of data objects which will be outsourced to the cloud server for clustering after encryption, and the cloud server perform k-means clustering are directly over the encrypted dataset without decryption. The client received intermediate results from cloud server.
8. PERFORMANCE EVALUATION

EXPERIMENTAL EVALUATION

1) **Experiment Configuration:** We evaluate the performance of our privacy-preserving MapReduce based k-means clustering scheme in terms of efficiency measured such as using Scalability, and accuracy.

2) **System setup:** The system setup terms are used to login the system as split data into encrypted. The dataset encryption cost maybe increases linearly from 1.26s to 6.34s when we change the size of dataset from 1 million objects to 5 million objects.

3) **Efficiency:** We focus on the efficiency of a single round clustering, because the different rounds of clustering’s have the same computational cost, but the number of clustering rounds is basically determined by the dataset itself and the selection of initial clustering centers.

4) **Scalability:** Evaluate the scalability of our scheme with respect to “scaleup”. Specifically, scaleup is one of the ability of using m-times larger resources to perform a m-times larger job in the same running time as the original job.

5) **Accuracy:** Compared with the original K-means clustering algorithms, in our scheme does not introduce any accuracy loss if all initial clustering centers are selected in the same way. In
particular, the allocation of a data object to the closest center is determined by the Euclidean distance between the object and the center.

9. IMPLEMENTATION

DATA PREPROCESSING MODULE

The figure 4 shows the data preprocessing module. First, the data set for employee management system is created which contains a set of tables such that employee details, salary details and work details for last four years, and this data first provide in MySQL database with help of this dataset we analysis this project.

![Figure 4 SQL Database](image)

**Experimental Results**

The figure 5 shows the MYSQL queries using csv document from excel sheet, with the data retrieved from the database and send it to HDFS in the Hadoop. The project is run on the vm workstation then the view employee details in rows and columns into the tables of itemset.
10. CONCLUSION

This concept successfully reveals the privacy secure MapReduce based on k-means clustering using employee datasets are fetch into the encrypted dataset, which means user sends the data in cloud computing by allocating some secrets encryption key used by light weight encryption key, this scheme achieves the goal of clustering speed and accuracy. The support of large-scale dataset, we securely integrated MapReduce framework into our design, and make it extremely suitable for parallelized processing in cloud computing environment. The security examination and numerical investigation are complete the execution of our plan in terms of security and effectiveness.

REFERENCES


