K-Means Clustering Over Map-Reduce For Data Privacy

- **Dhawala Mali , Latika Umrikar , Priyanka Panat, Nishita Tagadpallewar, Prof. R. S. Apare**

[1]*Smt. Kashibai Navale College of Engineering, Vadgaon(Bk)*
[2]*Smt. Kashibai Navale College of Engineering, Vadgaon(Bk)*
[3]*Smt. Kashibai Navale College of Engineering, Vadgaon(Bk)*
[4]*Smt. Kashibai Navale College of Engineering, Vadgaon(Bk)*

Abstract:

In data mining, for analysis of large dataset we used clustering technique. Clustering is the process of making a group of abstract objects into classes of similar objects. To manage this large data efficiently, we used cloud infrastructure for store those data and to perform clustering on those data. Since the data is sensitive in nature, the data stored on cloud should be secure so that we store encrypted data on the cloud. We propose a practical privacy preserving map reduce based k-means clustering scheme. In this scheme, the cloud server performs the clustering on encrypted datasets. We perform clustering using k-means clustering technique. We also investigate secure integration of Map Reduce into our scheme, because there is a large amount of data generation in the industry so its hard to deal with such big amount of data. To handle this data we are going to use map Reduce. It contains mappers and reducers to manage the data. This makes our scheme extremely suitable for cloud computing environment. This scheme gives the security of data and efficiency in performance. And makes suitable to the cloud environment. However dataset which require for the k-means clustering is very sensitive so that it has to be securely stored on cloud for further processing. i.e., patient data. Health data, behavioural data. Clustering is require to save your data on different clusters so that it's easy to get the data.

Keywords - *Privacy-preserving, K-means Clustering, Cloud Computing*

Introduction:

In the recent years data is generating day by day. There is a need to store this data somewhere where you will get high security to your data. The data should stored on cloud. Here we are combing cloud computing with data mining. In data mining there is a concept called clustering. In that we can use many algorithms like k-means, knn etc. In our project we used the concept of clustering with map reduce. Map reduce combine the output into compressed format. Why we use map reduce? Mp reduce is the tool of hadoop it is use to map your output and reduce it and give final result. The data can be sensitive in nature so here you can take example of patient's health record. So here there is lack of computation time. Except privacy protection, there are two different factors those are clustering efficiency and clustering accuracy. So the computational cost of the dataset owner will be reduced. There is no existing k-means clustering map reduce sign so we are developing a model to achieve comparable efficiency and accuracy to the clustering over unprotected data.

In this work, we proposed a privacy preserving k-means clustering technique for large amount of dataset. This can be outsourced to public cloud server. Depending on encryption technique we construct a full k-means clustering process based on learn with error hard problem. So it provides privacy preserving similarity checking directly on data objects called cipher text blocks in which cloud servers only have access to encrypted data.

Problem Definition:

Size of data has been increasing day-by-day, which requires a highly efficient system to analyze this data. We use the Clustering technique to analyze the data. In order to increase the efficiency of computation efforts have been made that there is an increasing interest in uploading data on public cloud. Since data is sensitive in nature and public cloud introduces privacy concern, the privacy of data should be considered.

Objectives

- To efficiently outsource large scale datasets to public cloud servers.
- To perform clustering directly over encrypted datasets,
- To improve the clustering performance in cloud computing environment
- To provide privacy of data and protect original data to be accessed from cloud provider.

System Architecture

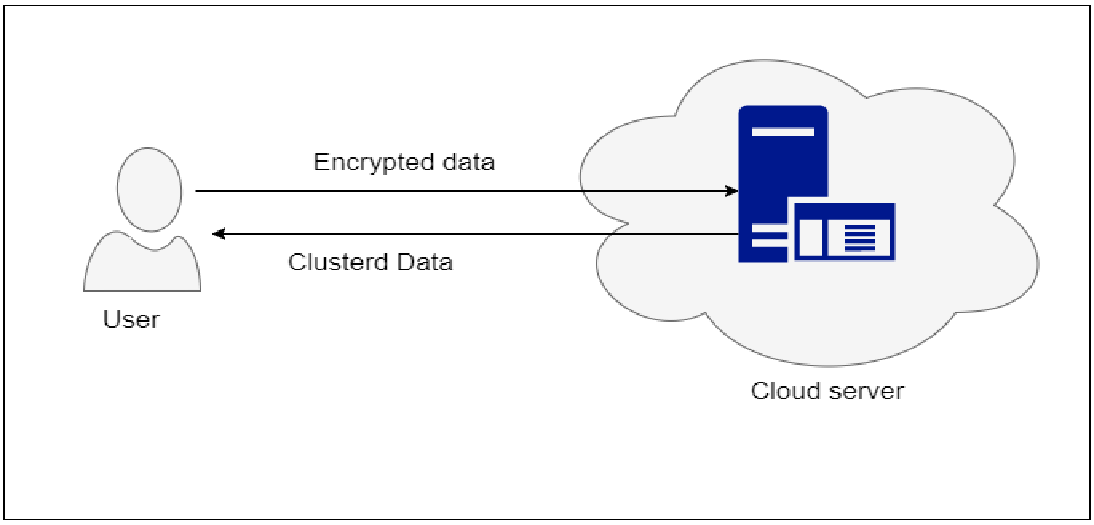

Fig 1: System Overview

In the above system overview the data is uploaded by the user. Here there is a k-means clustering with Map-reduce. Hadoop is used in this project. First encryption is performed on file and then there is k-Means clustering. In k-means clustering user has to form cluster by using centroid. This process will be continue till there is no same cluster formation from the centroid.

Related Works

1) Privacy-preserving data Mining.:

Year: 2000

Author Name:

- Rakesh Agrawal
- Ramakrishnan Srikant.

Description: To accurately estimate the distribution of original data values.

Limitations: Failed to achieve enough accuracy.

2) Revisiting Privacy Preserving Clustering by Data Transformation:

Year: 2003

Author Name:

- Stanley R. M.

- Oliveira
- Osmar R. Zaane.

Description: To address privacy-preserving clustering, in scenarios where data owners must not only meet privacy requirements but also guarantee valid clustering results.

Limitations: Do not achieve enough privacy.

3) An attacker's view of distance preserving maps for privacy preserving data mining:

Year: 2006

Author Name:

- Kun Liu
- Chris Giannella,
- HillolKargupta.

Description: To examine the effectiveness of distance preserving transformations in privacy preserving data mining.

Limitations: Adversaries who get a few unencrypted data records in the dataset will be able to recover rest records protected by data transformation.

4) Secure knn computation on encrypted databases:

Year: 2009

Author Name:

- Wai Kit Wong
- David Wai-lok Cheung
- Ben Kao
- Nikos Mamoulis

Description: To achieve privacy-preserving K-means clustering is to extend existing privacy-preserving K-nearest neighbors (KNN).

Limitations: Schemes are limited by the vulnerability to linear analysis attacks.

5) Parallel k-means clustering based on mapreduce.:

Year: 2009

Author Name:

- Weizhong Zhao
- Huifang Ma
- Qing He.

Description: To make the clustering method applicable to large scale data.

Limitations: Do not consider privacy protection for the outsourced dataset.

6) Secure nearest neighbor revisited.:

Year: 2013

Author Name:

- J. Byun
- H. Rhee
- H. Park
- D. Lee

Description: To enable the client to perform NN queries without letting the server learn contents about the query (and its result) or the tuples in the database.

Limitations: Support up to two dimension data

7) Privacy of outsourced k-means clustering.

Year: 2014

Author Name:

- Dongxi Liu
- Elisa Bertino
- Xun Yi

Description: The outsourcingof K-means clustering by utilizing homomophic encryption.

Limitations: The homomophic encryption utilized is not secure.

8) Optimized big data k-means clustering using mapreduce. :

Year: 2014

Author Name:

- W. Yau
- R. Phan
- S. Heng
- B. Goi,

Description: To address the problems of processing large-scale data Using K-means clustering algorithm.

Limitations: Do not consider privacy protection for the outsourced dataset.

Limitation of Study:

The storage overhead is introduced in the encryption of dataset and clustering centers. In our scheme, each data object and clustering center are denoted as a m-dimensional vector, and will be encrypted as two 2mdimensionalvectors. Thus, the total storage cost in our scheme is four times to that of the unprotected clustering.

Design of the Study

- Input: File in the form of text/Doc.
- Output: File
- Functions :

1 Upload the file in the encrypted form.

2 File gets uploaded on cloud.

3 Encryption is performed on data.

4 K-means clustering is performed on the data by considering the centroid.

5 That input is give to the map reduce framework.

6 Data is mapped and reduced to save the space.

- Success Conditions: File in clustered format.
- Failure Conditions: Storage overhead.

Tools Used

- **Software Requirement:**
  - Operating System : windows 8 and above.

○	Application Server	: Tomcat5.0/6.X
○	Language	: Java
○	Front End	: HTML, JSP
○	Database	: MySQL

- **Hardware Requirement:**

○	Processor	: Intel i3/i5/i7
○	RAM	: 4 GB (min)
○	Hard Disk	: 20 GB(min)

Statistical Technique Used

We have developed Login and Registration which manages the user profiles (User), here user registers and logins to the system. Then the encryption is performed on the data. This data is give as a input to k-means clustering so clustering is performed on the data and the data is uploaded on the cloud into chunks.

Algorithm

- **K-Means**: This algorithm is used to form a cluster using centroid.
- **Advanced Encryption Standard**: In our system, we have used DES to provide encryption to the uploaded files. Along with that we will be using conjunctive keyword to download the file.

Algorithm 1: Advanced Encryption Standard:

```
Cipher(byte in[16], byte out[16], key_array round_key[Nr+1])
begin
byte state[16];
state = in;
AddRoundKey(state, round_key[0]);
for i = 1 to Nr-1 stepsize 1 do
SubBytes(state);
ShiftRows(state);
MixColumns(state);
AddRoundKey(state, round_key[i]);
```

```
end for
SubBytes(state);
ShiftRows(state);
AddRoundKey(state, round_key[Nr]);
end
```

Algorithm 2: K-means Clustering

```
Input: k: number of clusters; max: a predefined number of iterations; n data objects
Output: k clusters
1: Select k initial cluster center such that 1 <= x <= k
2: for all predefined number of iterations
        1. Assign each data object to the cluster center with minimum distance it.
        2. Update cluster center to the average value of those data object assigned to the cluster x.
3: Output k reallocated clusters.
```

Our Approach:

The system will work in one operating modes:

1. User:

In this module user has to upload file. After that it will go to the encryption mode. And then the file gets divided into chunks and then there is k-means clustering performed and then the file is uploaded to cloud.

Experiment Result:

This system will perform k-means clustering using mapreduce framework of hadoop. The data will be clustered and you will get output files using using different centroid.

Future scope:

In future work, we are interested in how to reduce storage cost of clustering centers. In the project we used clustering to separate the data into different clusters and to save storage cost. And we are providing clustering on encrypted data to save data on public cloud and maintain the security because the data can contain anything like personal information and private data. So to perform this we are using clustering. We also investigate secure integration of MapReduce into our scheme, which makes our scheme extremely suitable for cloud computing environment. Thorough security analysis and numerical analysis carry out the performance of our scheme in terms of security and efficiency.

Acknowledgment: (optional)

It gives us great pleasure in presenting the preliminary project report on ‘K-Means Clustering over Map-Reduce for Data Privacy’.

I would like to take this opportunity to thank my internal guide Prof. R. S. Apare for giving me all the help and guidance I needed I am really grateful to them for their kind support. Their valuable suggestions were very helpful.

I am also grateful to Prof. B. H. Borhade, Head of Computer Engineering Department, *Smt. Kashibai Navale College of Engineering, Vadgaon(Bk)*, for his indispensable support and suggestions.

Dhawala Mali
Latika Umrikar
Priyanka Panat
Nishita Tagadpallewar
(B.E. Computer Engineering).

Conclusion:

This paper proposes a privacy preserved MapReduce framework with k-means clustering in cloud computing. This paper focuses only on security related to large-scale dataset. For that some techniques like encryption based on LWE hard problem used for fast accessing and accuracy without k-means clustering are being used. To manage large scale dataset the Map Reduce framework are being used because of security and large-scale dataset processing problem.

Reference:

1. Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. SIGMOD Rec., 29(2):439–450, May 2000.
2. Kun Liu, Chris Giannella, and Hillol Kargupta. An attacker’s view of distance preserving maps for privacy preserving data mining. In Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases, PKDD’06, pages 297–308, Berlin, Heidelberg, 2006. Springer-Verlag.
3. Dongxi Liu, Elisa Bertino, and Xun Yi. Privacy of outsourced k-means clustering. In Proceedings of the 9th ACM Symposium on Information, Computer and

Communications Security, ASIA CCS ’14, pages 123– 134, New York, NY, USA, 2014. ACM.

4. Wai Kit Wong, David Wai-lok Cheung, Ben Kao, and Nikos Mamoulis. Secure knn computation on encrypted databases. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, SIGMOD ’09, pages 139–152, New York, NY, USA, 2009. ACM
5. Weizhong Zhao, Huifang Ma, and Qing He. Parallel k-means clustering based on mapreduce. In Proceedings of the 1st International Conference on Cloud Computing, CloudCom ’09, pages 674–679, Berlin, Heidelberg, 2009. Springer-Verlag.
6. Jiawei Yuan and Shucheng Yu. Privacy preserving back-propagation neural network learning made practical with cloud computing. IEEE Transactions on Parallel and Distributed Systems, 25(1):212–221, 2014.
7. Zvika Brakerski, Craig Gentry, and Shai Halevi. Packed ciphertexts in lwe-based homomorphic encryption. In 16th International Conference on Practice and Theory in Public-Key Cryptography (PKC), pages 1–13, February 2013.