JOURNAL FOR INNOVATIVE DEVELOPMENT IN PHARMACEUTICAL AND TECHNICAL SCIENCE

Volume-1,Issue-1 (Oct-2018)
ISSN (O) :- 2581-6934

Implementation of Map Reduce Based Clustering for Large Database in Cloud

P.NARESH[1], G.VENU BABU[2], S KRUPAMAI YENDRAPATI[3], K.GURNADHA GUPTA[4],M.KIRAN KUMAR[5]

[1]Research Scholar, VTU, Chennai ,[2]Asst.Professor, SICET, Hyderabad ,[3] Asst.Professor, BWEC, Bapatla
[4, 5] Research Scholar, SSSUTMS, Bhopal,MP,INDIA

Abstract- Nowadays clustering takes a crucial role in data mining. Clustering stands for grouping similar data items into a single place. Inside the cluster there is high intra cluster similarity. In real world Social sites data, marketing, banking and industrial data as going on increasing day by day. To handle those data is also a major problem which involves some privacy issues. All existing systems use normal clustering techniques that can reduces the mining process but there is lack of sensitive information and effort to mine data is complex. So the proposed MapReduce based privacy preserving techniques will efficiently outsource the data into cloud and clustering operations performed on data being encrypted. Compared with existing methods, proposed framework will give accurate results and suitable for cloud data.

Keywords – Cluster, MapReduce

1. INTRODUCTION

Clustering is a standard procedure in multivariate data analysis. it's designed to explore AN inherent natural structure of the data objects, where objects inside identical cluster ar as similar as possible and objects in various clusters ar as dissimilar as possible. The equivalence categories elicited by the clusters give a method for generalizing over the info objects and their options. cluster strategies ar applied in several domains, like medical analysis, psychology, political economyand pattern recognition.Clustering is an exploratory data analysis.Therefore, the someone might need no of very little data concerning the parameters of the ensuing cluster analysis. In typical uses of

Clustering the goal is to determine all of the following:

- The number of clusters,
- The absolute and relative positions of the clusters,
- The size of the clusters,
- The shape of the clusters,
- The density of the clusters.

The k-Means clustering algorithm is an unsupervised hard clustering method which assigns the n data objects 0,1.. on to a pre-defined number of exactly k clusters C1,… Ck. The optimizing criterion in the clustering process is the sum-of-squared-error E between the objects in the clusters and their respective cluster centroids cen1,…cenk, K-means clustering is an iterative process that requires the update of clustering centers based on the entire dataset after each round of clustering. Considering the efficient support over large-scale datasets, these update processes also need to be outsourced to the cloud server in a privacy-preserving manner. There are so many MapReduce clustering methods are present but all are lack of providing security for outsourced data in cloud. In this work, we proposed a practical privacy-preserving K-means clustering scheme for large-scale datasets, which can be efficiently outsourced to public cloud servers. Our proposed scheme simultaneously meets the privacy, efficiency, and accuracy.

2. Background Knowledge

In the below diagram there are two major components are there. One is Cloud server and another one is Dataowner. In

that dataowner having set of data objects which will outsourced to cloud server for providing security via encryption at the same time Clustering also done on given data set at server side. While the clustering operation is going on, the cloud server interacts with the dataowner for inputs. Those input data given by dataowner is helpful in clustering.

The cloud server having access on encrypted data which is generated by server. In background model the server of cloud having additional information related dataset. The server was unable to fetch the cluster centroids. In this way the proposed

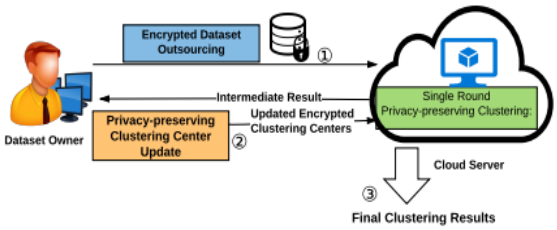


Fig.1.Architecture

architecture maintains privacy and security for data apart from server and outsiders.

3. The Model and Preliminaries

K-Means: The main purpose of k-means algorithm is to cluster similar data objects of same type into a cluster. It relocates data objects into different clusters depending on object weight and centriod based ecludien distance. The objects which are irrelevant are left outside the cluster called outliers. Highintra similarity inside the cluster and high inter similarity outside the cluster will be maintained while performing clustering.

The below steps illustrate about k-means clustering.

Given a set $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{R}^m$

a distance measure d on $\mathcal{R}^m$

a function for computing the mean $\mu : \mathcal{P}(\mathcal{R}^m) \to \mathcal{R}^m$

Select (arbitrarily) k initial centers $f_1, \ldots, f_k$ in $\mathcal{R}^m$

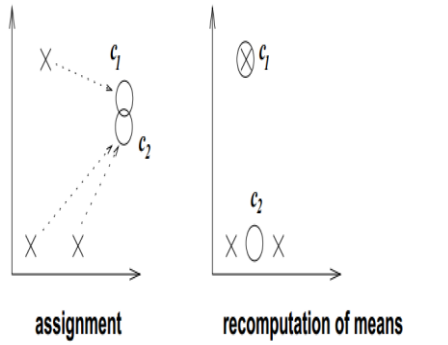
while the stopping criterion is not true

for all clusters c_j do $c_j = \{x_i \mid \forall f_l \; d(x_i, f_j) \leq d(x_i, f_l)\}$ end

for all means f_j do $f_j \leftarrow \mu_j$ end

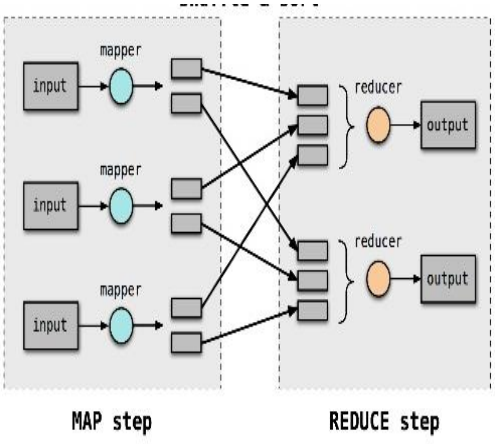
end

Illustrating the k-Means Clustering Algorithm



Along with k-means clustering,MapReduce method applied on large dataset. In MapReduce two parts are there. One is map function and another is Reduce function. It divides dataset into small pieces which are easy to process. Map function process data and produce intermediate output as <key,value> format. These results further forwarded to reduce function to do reduce operation and review all outputs and gives final result

MapReduce Architecture:



ISSN:-2581-6934|www.jidps.com

Encryption for security: To provide security and privacy there should be Data Encryption algorithms were used, which are Map privacy and Reduce Privacy steps.

4. Conclusion

The proposed MapReduce based K-means clustering scheme in cloud computing is exercised on datasets. Our scheme achieves clustering speed and accuracy that are comparable to the K-means clustering without privacy protection. The frame work which was proposed is suitable for large datasets in cloud which yields good results by means of accuracy and time. Compared with existing methods, proposed framework will give accurate results and suitable for cloud data.

References

[1]. Jiawei Yuan, Membe-IEEE, YifanTian, Student Member-IEEE, "Practical Privacy-Preserving MapReduce Based K-means clustering over Large-scale Dataset" 2016 IEEE.

[2].JaideepVaidya and Chris Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, pages 206–215, New York, NY, USA, 2003. ACM.

[3]. GeethaJagannathan and Rebecca N. Wright. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In Proceedings of the Eleventh ACM SIGKDD International Conference on
Knowledge Discovery in Data Mining, KDD '05, pages 593–599, New York, NY, USA, 2005. ACM.

[4]. Paul Bunn and RafailOstrovsky. Secure two-party k-means clustering. In Proceedings of the 14th ACM Conference on Computer and Communications Security, CCS '07, pages 486–497, New York, NY, USA, 2007. ACM.

[5]. Mahir Can Doganay, Thomas B. Pedersen, Y ¨ ucelSaygin, ErkaySavas¸and Albert Levi. Distributed privacy preserving k-means clustering with additive secret sharing. In Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, PAIS '08, pages 3–11, New York, NY, USA, 2008. ACM

[6]. Jun Sakuma and Shigenobu Kobayashi. Large-scale k-means clustering with user-centric privacy-preservation. Knowledge and Information Systems, 25(2):253–279, 2009.

[7]. Xun Yi and Yanchun Zhang. Equally contributory privacy-preserving k-means clustering over vertically partitioned data. Inf. Syst., 38(1):97–107, March 2013.

[8]. RakeshAgrawal and RamakrishnanSrikant. Privacy-preserving data mining. SIGMOD Rec., 29(2):439–450, May 2000.

[9]. B.M.G. Prasad, P. Naresh, V. Veeresh, "Frequent Temporal Patterns Mining With Relative Intervals", International Refereed Journal of Engineering and Science ,Volume 4, Issue 6 (June 2015), PP.153-156.

[10]. Stanley R. M. Oliveira and Osmar R. Zaane. Privacy preserving clustering by data transformation.In Brazilian Symposium on Databases, SBBD, Manaus, Amazonas, Brazil, 2003.

[11]. Kun Liu, Chris Giannella, and HillolKargupta. 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

[12]. H. Kargupta, S. Datta, Q. Wang, and KrishnamoorthySivakumar. On the privacy preserving properties of random data perturbation techniques.In Data Mining, 2003.ICDM 2003. Third IEEE International Conference on, pages 99–106, Nov 2003.

[13]. 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.

[14]. Sen Su, YipingTeng, Xiang Cheng, Yulong Wang, and Guoliang Li. Privacy-preserving top-k spatial keyword queries over outsourced database. In Proceedings of the 20th International Conference on
Database Systems for Advanced Applications, DASFAA'15, pages 589–608, and 2015.

[15]. 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.