INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 1)
Available online at www.ijariit.com

Security and Clustering Of Big Data in Map Reduce Framework: A Survey

Anju Abraham
anju.anuja08@gmail.com
Musaliar College of Engineering and Technology, Kerala

Shyma Kareem
shymshiju@gmail.com
Musaliar College of Engineering and Technology, Kerala

ABSTRACT

For maintaining the authenticity, privacy and confidentiality of larger dataset are outsourced to the cloud in the encrypted format. Cloud storage provides data management and reduces the costs. Various clustering methods like K mean, K nearest niebouring, DBSCAN clustering methods are implemented to cluster massive data that are related to each other using map reduce framework in bigdata analytics. Clustering was done on encrypted partitioned data in order to protect the information from the third party access. Various approaches have been used for securing and maintaining the efficiency and performance of millions of dataset with variety, velocity, and volume.

Keywords: ***K- mean Clustering, Privacy, K- nn Clustering, Partitioned Data, Map Reduce.***

1. INTRODUCTION

As cloud computing is offering various services like scalability, flexibility and larger data management with low maintenance cost the various organization, institutions and companies in different sector prefer to outsource there data to the cloud platform. While outsourcing data to the cloud must ensure confidentiality, privacy, and security of the data. Dataset consists of sensitive information like personal health information, localized data, financial data and personal photos etc. which should not be disclosed to the third party. On one hand data owner encrypt the data before launching it to the cloud environment. On the other hand, several schemes are used to securely query the data in the cloud storage by preserving the privacy of the dataset. With the rapid growth of big data in its variety, velocity and volume clustering methods are implemented that adheres to these features. For example to predict the investment of person requires clustering over his financial records, his savings etc. that contains sensitive information. In the multi-party scheme, cryptographic primitives are used that are expensive as it includes homomorphism encryption and transfer of data. Another research targeted on privacy-preserving clustering based on distance metrics. Various query processing schemes are used by sharing or not sharing the keys by the data owner. Clustering like knn clustering, K-mean clustering, and DBSCAN clustering has been implemented. Knn uses a single round of search while k-mean uses iterative approaches and DBSCAN is used for the larger dataset. Research has been carried out to deal with the large dataset in parallel.

2. MAP REDUCE FRAMEWORK

Hadoop Map Reduce is the software paradigm for processing larger massive and scalable dataset in the cluster. Map Reduce model processes basically the unstructured dataset available in a clustering format. As the name indicates it mainly has two jobs map and reduce job. It splits the input data into smaller chunks and processes it in parallel. A map operation takes a set of input data and converts it into an intermediate <key, value> pair. The Reduce operation takes those intermediate <key, value> pair and sorts it with the matching key to give the final output.

The Reducer has three main phases:

- Shuffle
- Sort
- Reduce

Shuffle Phase input to the reducer is the output obtained from the mapper. Different map operation produces different output thus the framework fetches the required portion of the mapper by using HTTP. In the sort phase values are grouped according to the key. Different mapper has the same key. Shuffle and sort process is carried out in parallel. Reduce Phase calls the reduce method that takes <key, list of corresponding value> pair and produces the output into the file system. Most of the security is provided in the map and reduce framework in order to protect the numerous amount of sensitive dataset.

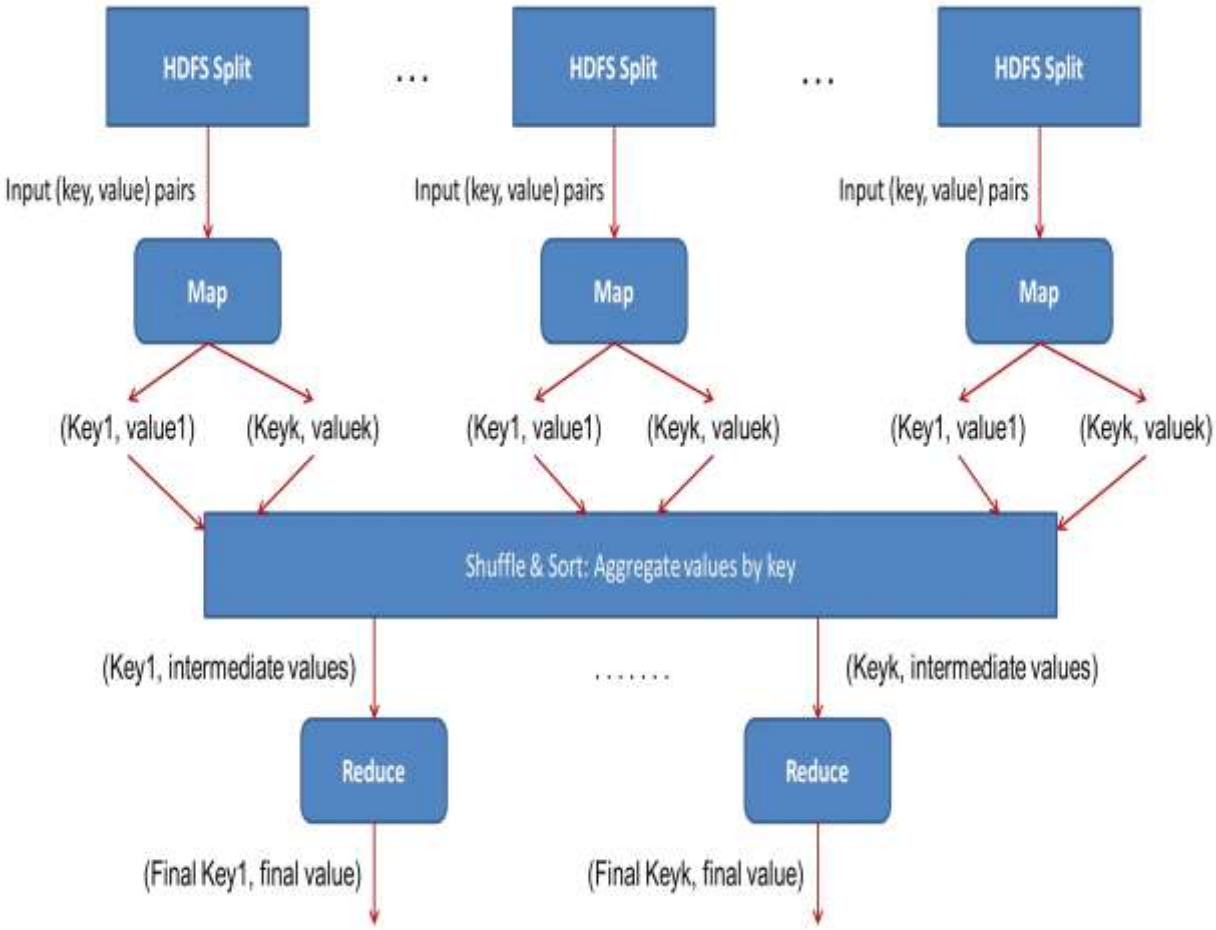

Fig 1: Map Reduce Framework

3. CLUSTERING AND PRIVACY

In most of the research work cryptographic algorithms are carried out in the clustered data and clustering is done on the object available in the cloud or in Hadoop distributed file system. This is done in order to provide privacy of the dataset from the external intruders. Some of the researchers have proposed new clustering techniques while others have used the existing approaches with some enhancements in order to provide efficient and accurate clusters with minimal outliers. Clustering is the unsupervised technique that does not use any class labels and thus groups larger dataset based on different clustering metrics. All the related dataset are grouped together that are separated from the unrelated set. This forms the cluster. This paper surveys clustering in a larger dataset. Clustering is carried out keeping in mind the different factors like size of the data, data can be in terabytes, pent bytes etc., type of the data like numerical data, binary data, ordinal data and categorical data etc, shape of the cluster spherical, arbitrary shape etc, time complexity and dealing with the outliers, number of clusters required, it is basically defined by the user before running the algorithm.

Choice of objects that need to be clustered, dealing with missing data in the databases for this various technique like regression method, smoothing or binning methods are used to overcome the same, dissimilarity measures and metrics to identify the similar object. Various clustering methods carried out by the researchers on big data along with the cryptographic algorithms are as follows:-

A. Partitioning Methods

In partitioning method the data objects that are similar to each other are categorized into the same partition, the clusters thus obtained do not overlap with each other. The clustering algorithms are k-mean weighted k-mean, k-medoids, k median etc. In k-mean algorithm is an iterative approach in each iteration it finds new cluster head and reassigns each of the objects to the new cluster head based on its closeness. The closeness is estimated by using the Euclidian distance.
The termination criterion is based on the squared error. In k-mean, the mean of the data object that belongs to a particular cluster is

calculated. In weighted k-mean, each object is assigned with the weights based on its importance. In k-medoids and median, the medoids and the median of the data object are obtained. Other partitioning methods used are CLARA and CLARANS for the larger dataset.

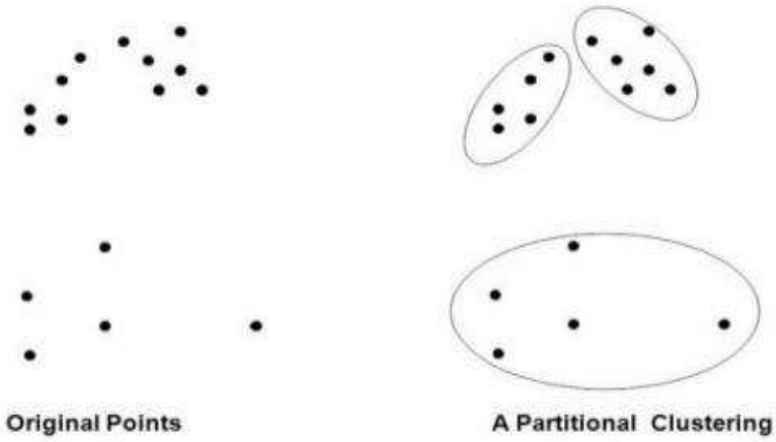

Fig.2 Partitioning Clustering

Advantage
1. Handles larger dataset efficiently.
2. Easier implementation.
3. It is faster and produce dense cluster.

Disadvantage
1. Do not handle noisy data and outliers.
2. It generates empty clusters.
3. Not suitable non-spherical cluster formation.

B. Hierarchical Methods

Hierarchical clustering method forms the tree like clusters in the form of nested clusters. Clusters are formed either recursively or by iteratively partitioning the dataset. Two main approaches used for grouping of the data objects are top down and bottom up approaches. Bottom up approach is agglomerative that begins by placing each object as an individual cluster and merging similar cluster to obtain a new cluster, this continues till stopping criterion is achieved or a suitable cluster is obtained. Divisive is to down approach that is just the reverse of agglomerative. In divisive approach enter cluster is considered as a single cluster, it is then further divided to obtain an individual cluster. Merging or splitting is carried out based on the criterion like a single link, complete link, average link etc. Examples of hierarchical clustering are BRICH, CURE, and chameleon used for the larger dataset or for the dataset representing higher dimensionality.

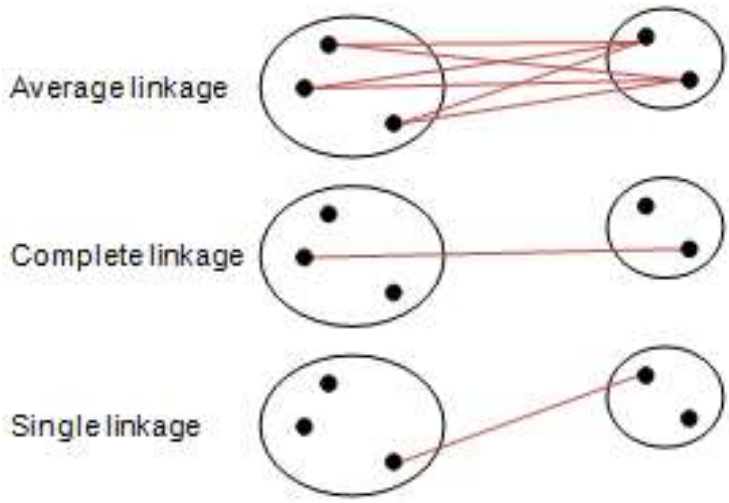

Fig 3. Hierarchical Clustering

Advantage
1. It is less sensitive to outliers and unwanted data.
2. It forms n number of clusters.
3. It can be implemented to a dataset of any type.

Disadvantage
1. Once splitting and merging are done it cannot be undone to improve the cluster quality.
2. It does not scale efficiently.

Other clustering methods that can be implemented are density based clustering, grid based clustering and model based clustering etc.

C. Privacy on Clusters

Most of the research work is based on attribute based encryption method and fully homomorphic based encryption scheme. In both the scheme the data owner ensures that clustering should be done on the encrypted data object decryption is also done by the data owner if required. Different variations and enhancement are done on the same or a new scheme has been implemented. In attribute based encryption it uses two policy: cipher text based and key based policy in cipher text based the owner of the data controls the access of data, based on the complexity the design of the public key is made more difficult. Thus researches concentrate more on the design of access strategy. In case of policy based a set of attribute list is used to obtain encrypted text and private key.

In fully homomorphic allows performing computation on the encrypted data. It performs various mathematical operations without any decryption of the dataset. This thus reduces the complexity of decryption and encryption which otherwise would have to be carried out frequently. Searchable encryption scheme offers a secure searching function to search the encrypted data whenever the query is generated by the user. Other encryption methods used are a hash based method and various other algorithms based on the type of data.

4. LITERATURE SURVEY

In this section, we highlight the various privacy protection methods carried out as part of research work as shown in the table. In the recent years, many of the research work has been proposed to secure and provide privacy to the big data. Privacy is provided by adopting novel cryptographic techniques to the outsourced data or to the data maintained in Hadoop.

TABLE 1: PRIVACY AND CLUSTERING

Sl no.	Title	Problem Statement	Merits	Demerits
1.	Top-K Spatial Keyword Queries over Outsourced Databases.	Encrypted tree index is build to facilitate security in outsourced spatial data.	Provided scalability and efficiency.	Caused plaintext attack and concentrated on spatial data.
2.	Privacy of outsourced K-mean Clustering.	This paper proposed avoiding of secure division operation that was used in computing cluster centre.	It provided a comprehensive solution and protect data confidentiality.	Faced scalability and performance issues.
3.	Optimized Big Data Clustering using Map Reduce.	Processing large scale data using clustering algorithm and proposed novel processing model in map reduce.	It was robust and scalable.	Had security related issues
4.	Equally Contributory Privacy Preserving Clustering Over Vertically Partitioned Data.	Multi-party clustering for vertically partitioned data.	Clustering data without telling intermediate search results.	Scalability issues and was feasible only for the small dataset and could not support larger dataset.
5.	Fully Homomorphic Encryption Schemes Without Bootstrapping.	Secure circuit evaluation using homomorphic encryption to achieve collaborative secure computation among multiple parties.	Better security in a collaborative environment.	It is inefficient for large scale datasets and is costly.
6.	Efficient Privacy Preserving Biometric Identification in Cloud Computing.	Identifying client in a cloud environment using biometric traits-fingerprint based identification is deployed.	Client fingerprint is not disclosed to a cloud platform.	Laborious computation is required.

7.	Preserving Privacy in Data Mining.	Providing various privacy models to protect against de-anonymised the target records.	Takes into consideration the sensitive information and attribute values. Owners are allowed to define their privacy levels as per their requirements.	Still blocking certain queries reveals some of the important information.
8.	Distributed Privacy-Preserving Clustering with Secret Key Sharing.	Novel Protocol is Proposed for Preserving Privacy in Clustering based on secret sharing of keys.	Data attributes are maintained in different sites thus providing better computation cost and parallel processing.	Required support for terabytes and pent bytes of data.
9.	Privacy Preserving for Arbitrarily Partitioned data.	A generalized version of both horizontal and vertical partitioned data and providing privacy protocol for the clusters.	Allow communication without revealing individual and personal data items.	Not scalable.

5. CONCLUSION

In this paper, we have discussed various privacy protection and security mechanism and methods which can be used for security purposes in the cloud platform. We have also surveyed the clustering methods and algorithms implemented for grouping the larger dataset. We also highlighted on map reduce framework. These mentioned methods conclude that there are lots of algorithms and techniques used in cloud platform using map reduce framework to preserve the privacy and authenticity of information and it also needs lots of improvisation due to enabling technology.

REFERENCE

[1] Geetha Jagannathan and Rebecca N. Wright, "Privacy-preserving dis-tributed k-means clustering over arbitrarily partitioned data," ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, pages 593–599, New York, NY, USA, 2005.

[2] Paul Bunn and Rafail Ostrovsky, "Secure two-party k-means clustering," ACM Conference on Computer and Com-munications Security, CCS '07, pages 486–497, New York, NY, USA, 2007.

[3] Mahir Can Doganay, Thomas B. Pedersen, Yucel¨ Saygin, Erkay Savas¸, and Albert Levi, "Distributed privacy preserving k-means clustering with additive secret sharing," International Workshop on Privacy and Anonymity in Information Society, PAIS '08, pages 3–11, New York, NY, USA, 2008.

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

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

[6] Dongxi Liu, Elisa Bertino, and Xun Yi, " Privacy of outsourced k-means clustering," the 9th ACM Symposium on Information, Computer and Communications Security, ASIA CCS '14, pages 123– 134, New York, NY, USA, 2014.

[7] Yongge Wang, "Notes on two fully homomorphic encryption schemes without bootstrapping," Cryptology ePrint Archive, Report 2015/519, 2015.

[8] B. Yao, F. Li, and X. Xiao, "Secure nearest neighbor revisited," Data Engineering (ICDE), 2013 IEEE 29th International Conference on, pages 733–744, April 2013.

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

[10] 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.

[11] Apache hadoop. http://hadoop.apache.org/.

[12] http://en.wikipedia.org/wiki/

