Secured Privacy Preserving Technique for Big Data on Cloud using High Order Clustering

M. S. Premalatha
Research Scholar, Manonmanium Sundaranar University, Abishekapatti, Thirunelveli – 12, Tamil Nadu, India.
premalatha_ms@yahoo.co.in

Dr. B. Ramakrishnan
Associate Professor, Department of Computer Science and Research Centre, S.T. Hindu College, Nagercoil, Tamil Nadu, India.
ramsthc@gmail.com

Abstract – In this research, a privacy-preserving high order great amount of heterogeneous information victimization distributed high order bunch algorithmic program is projected. The privacy-preserving high order cluster the heterogeneous information set by representing every heterogeneous data object as a tensor. In this paper, the cloud server directly performs bunch over encrypted datasets, whereas achieving more accuracy. The absolutely homomorphic coding algorithmic program (Fully Homomorphic Encryption) is used to safeguard the high order great amount of heterogeneous information. Moreover, a secure integration of map-reduce is meant into this projected work, that makes this work tremendously acceptable for cloud computing.

Index Terms – Privacy-Preserving, Heterogeneous Data, Clustering, Cloud Computing, Map Reduce, Encryption Algorithm.

1. INTRODUCTION

Recently, the expansion of data has developed immensely in their range, dimension, altering from centralized to the distributed atmosphere, and the period of big data. Big data is generally indicating the large or compound data sets with the capability of conventional data processing and technique. Additionally, Big data are heterogeneous where every entity in big data set is multi-representative [1]. Particularly, big data sets are encompassing diverse varieties of entities like texts, images, and audios as structured and unstructured data. Here, dissimilar kinds of entities are also used to manage the dissimilar information at the same time as they are consistent with each other [2]. Mainly, the appearance of big data sets is used to engender chances for forecaster, researchers, and business people. Also, it is permitting to incorporate big data to take choice through finding a hidden model, unidentified correlations, and supplementary approaches as systematic.

Here, the enormous information investigation is additionally fused with various complex information mining methodology like grouping [3]. The Facebook is considered as the noteworthy delineation for huge information investigation, where the prominent social sites are gathering 500 terabytes (TB) information normally [5]. In this way, this attribute is passed on a requesting issue to grouping advancements. For the most part, the bunching is considered as the premier errand of looking at information mining and arithmetical information examination, which is actualized in various zones like human services, interpersonal organization, picture investigation, design acknowledgment, and so on. These days, the snappy development of enormous information in information mining and examination is likewise showing up against bunching through volume, assorted variety, and speed [4]. Also, huge information commonly include incredible amount of information. Here, the protection upkeep is a chief issue for huge information mining applications. For the most part, the Privacy Preserving Data Publishing includes two portions, information accumulation and information distributing. In information accumulation section, the dataset is made by methods out of information distributer from information proprietor [6]. A short time later, the accumulation of the natural dataset occurs in the information distributing portion. Here, the prepared dataset is additionally transmitted to the information beneficiary. Job related access control is considered as a methodology which offers directions for the information by including confirmation or access control. In this way, the delicate information is essentially accessible for confined customer gatherings. The cryptographic system is an extra procedure. Here, the touchy information zones are anonymized for they can't recognize to element check [7].

Normally, the considered foremost confront of uncertainty in the data is facilitating a number of isolation objectives. Map-Reduce is an equivalent and disseminated large-scale data processing model which is broadly investigated and roughly implemented for big data applications [8]. Map-Reduce is a dominant, expandable, and commercial for the relevant uniqueness of cloud computing which is used to incorporate the communication possessions through cloud systems [9]. Generally, a Map-Reduce includes two primitive tasks such as Map and Reduce. The execution of Map task through

illustration is known as Mapper. On the other hand, the execution of Reduce task is known as Reducer. The foremost intention of big data clustering is constructing competent and effectual parallel clustering algorithms [10]. Here, the Map-Reduce process is mostly carried out by these algorithms. The foremost process of clustering is combining the objects through their resemblance. Therefore, the comparable objects are positioned in the identical group (cluster) and unrelated objects are positioned in dissimilar clusters [11]. In cloud systems, the presentation of large-scale datasets on the solitary computer is enhanced by means of Map-Reduce programming representation which takes place in the data cluster.

2. LITERATURE SURVEY

Clustering is a famous procedure for big data analytics and mining. On the other hand, several existing algorithms are not effectual to cluster heterogeneous data in big data. Therefore, Fanyu Bu et al. [12] have depicted an algorithm for clustering. Mainly, it is based on three processes such as (i) an adaptive dropout deep learning representation which is used to find out attributes from every kind of data, (ii) an attribute tensor representation and (iii) a tensor detachment-related high-order CFS algorithm.

Jiawei and Yifan [13] have portrayed a reasonable security shield K-implies bunching design which is skillfully subcontracted by cloud servers. This configuration is primarily encouraging cloud servers to do bunching as specifically in encoded datasets which are likewise used to achieve the computational intricacy and precision than the further grouping procedure. Also, it is utilized to look at safe consolidation of Map-Reduce as their organization, which develops their configuration astoundingly fitting for cloud computing environment. Here, the wellbeing examination and arithmetical examinations are executing the introduction of organization through resistance and adequacy. Additionally, Peng Li et al. [14] have delineated a protection safeguarding high-arrange neuro-fluffy c-implies calculation (PPHOFCM) which is mostly utilized bunching heterogeneous information in cloud computing. Here, the PPHOFCM is gathering the heterogeneous dataset through heterogeneous information object like a tensor and separation which are utilized to confine the connections through the high request tensor space.

The recognition of organization and investigation of big data is considered as the significant requirements for the absolute quantity and enhancing complication of data. Here, the fuzzy clustering algorithm is carried out the hard clustering methods through exactness. Simone Ludwig [15] has established an innovative policy to examine the parallelization and scalability of effectual algorithm such as FCM. The algorithm is usually parallelized by means of MapReduce model which is specifying the Map and Reduce primitives. The obtainable algorithm scales are finely expanded by dataset sizes. In cluster process, the datasets are encompassing responsive information which is openly subcontracting the public cloud servers as unavoidably expanding isolation concern. Jiawei Yuan and Yifan Tian [13] also have depicted a sensible privacy-defend K-means clustering format which is competently subcontracted to cloud servers. This format is mainly facilitating cloud servers to carry out clustering as openly in encrypted datasets. At the same time, it is also used to accomplish the equivalent computational complication and exactness than the further existing clustering process. Moreover, the Map-Reduce structure is firmly incorporated into the obtainable process. This process is mainly appropriate for cloud atmosphere.

3. SYSTEM MODEL

The background of the proposed privacy preserving big data clustering process is explained here. After the background explanation, the proposed part is explained. The fundamental aim of the proposed procedure is to structure a proficient protection safeguarding high request vast measure of heterogeneous information bunching in cloud computing system. The framework is given in figure 1. The framework includes two noteworthy units like dataset owner (DO) and cloud server (CS). The DO contains a lot of information object, which will be encoded and given to the cloud server for grouping. The cloud server plan with a map and reduce stage. The CS performs bunching to the encoded informational collection and the middle of the road results are passed to the DO. The DO unscramble the information utilizing the key. The grouping is finished when the bunching results don't change any longer, or a predefined number of cycles is obtained.

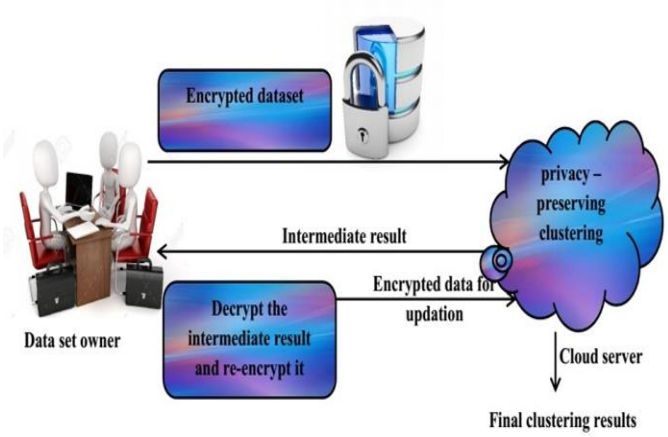

Figure 1. System Model

3.1. Big Data Analysis

Immense information is a word used for the portrayal of enormous measures of information which are sorted out, semi-composed or unstructured. The information, on the off chance that it can't be taken care of by the customary databases and software innovations at that point; such information is sorted

as large information. The enormous information is described using three V's.

- Volume: various segments add to the expansion in volume like stockpiling of information, live gushing and so forth.
- Variety: distinctive sorts of information are to be upheld.
- Velocity: the speed at which the documents are made and forms are done alludes to the speed.

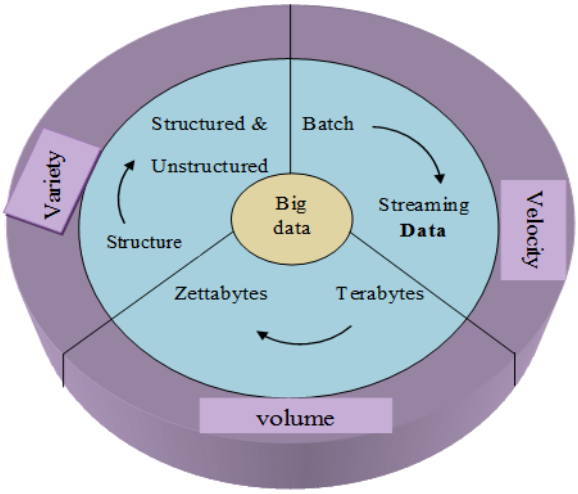

Figure 2. General Architecture – Big Data Analytics

In enormous information examination, map reduces undertaking is ordinarily utilized to reduce the information measurement. For instance, the computational multifaceted nature will increment altogether with the developing measure of heterogeneous information. The crucial course of action of huge information is indicated in figure 2.

3.2. Map-Reduce System

Map Reduce is a system which is for the most part utilized for handling a lot of information in a distributed way. To deal with the lot of information, Map Reduce separates the errand into two expressions, for example, Map and Reduce. The huge information is prepared in Map and reduce work which is called mapper and reducer. For enormous information handling, the map-reduce structure split the information into various gatherings and each gathering prepared by the autonomous mapper. At that point, each map work forms the information and deliver the yield as <key, value> sets. From that point forward, the yields are rearranged and given to the reducer expression. At that point, the reduce sort the yield and deliver the last outcome. The map-reduce structure is given in figure 3.

Information: Involved database

Map: separating and arranging the information

Decrease: redistribute the mapped information dependent on missing qualities and likeness

Output: All reduced information

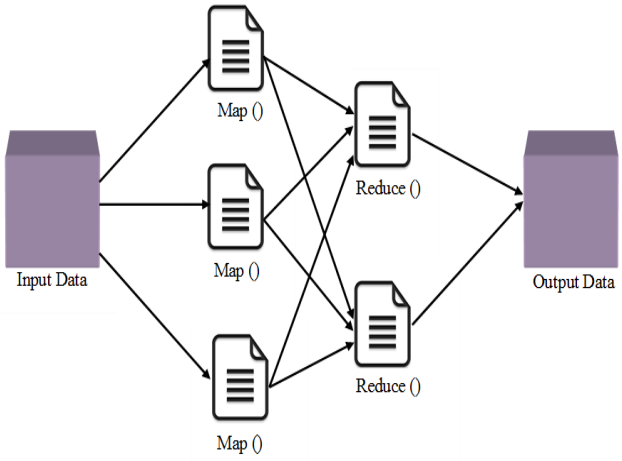

Figure 3. Map Reduce Framework

3.3. Construction of privacy-preserving high order clustering algorithm on cloud

The main objective of the proposed methodology is to design an efficient privacy-preserving high order large amount of heterogeneous data clustering in cloud computing network. At first, consider the heterogeneous dataset $Y = \{y_1, y_2, ..., y_n\}$ The aim of the privacy-preserving high order clustering is to group Y into c clusters by performing secure high order clustering algorithm on the cloud without disclose of private data. The proposed system mainly consists of three stages such as (i) system setup and data encryption, (ii) privacy preserving clustering on the Map-reduce framework and (iii) clustering center updation on the cloud. In the first stage, the data owners (DO) setup system parameters for clustering and Map Reduce. Then, DO encrypt the dataset for clustering using Fully Homomorphic Encryption (FHE) algorithm. In the second stage, the encrypted data is given to the cloud server for performing clustering. Here, in map phase, the encrypted objects are grouped to closest clustering center. Then in the reducer phase, centroids values are updated. Once the clustering process is completed, the cloud server sends the intermediate result to the data owner. After that, DO download the intermediate result and decrypt the data using an encryption key. Then re-encrypt the data and upload to the cloud server.

4. CONCLUSION

In this paper, a high-order bunch algorithmic program for heterogeneous information bunch is projected. Moreover, to boost the potency of cloud computing, there is a tendency to develop a privacy-preserving high-order bunch technique. The

absolutely homomorphic coding (FHE) algorithmic program is employed to safeguard the non-public information once playing the high-order bunch on the cloud. Moreover, the Map-Reduce framework is firmly integrated into the projected style to attenuate the difficulties of the large-scale dataset. The experimental results clearly show that the projected methodology achieves most bunch accuracy compared to different strategies. During this work, the projected algorithmic program is assessed on one delegated datasets.

REFERENCES

[1] B. Ermis, E. Acar, and A. T. Cemgil, "Link Prediction in Heterogeneous Data via Generalized Coupled Tensor Factorization," Data Mining and Knowledge Discovery, vol. 29, no. 1, pp. 203-236, 2015.

[2] Q. Zhang, L. T. Yang, and Z. Chen, "Deep Computation Model for Unsupervised Feature Learning on Big Data", IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 161-171, Jan. 2016.

[3] Zakaria Gheid and Yacine Challal, "Efficient and Privacy-Preserving k-means clustering For Big Data Mining", IEEE computer society, 2016.

[4] European Network and Information Security Agency. Cloud computing security risk assessment.https://www.enisa.europa.eu/activities/riskmanagement/files/deliverables/cloud-computing-risk-assessment.

[5] N. Soni and A. Ganatra, "MOiD (Multiple Objects Incremental DBSCAN)- A Paradigm Shift in Incremental DBSCAN", International Journal of computer science and information security, vol. 14, no. 4, pp. 316-346,2016.

[6] Data Mining With Big Data Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, And Wei Ding, Senior Member, IEEE, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No.1, January 2014.

[7] Review on Data Mining with Big Data" Vitthal Yenkar, Prof. Mahip Bartere, IJCSMC, Vol. 3, Issue. 4, April 2014.

[8] Dean and S. Ghemawat, "MapReduce: A Flexible Data Processing Tool," Communications of the ACM, vol. 53, no. 1, pp. 72-77, 2010.

[9] K.-H. Lee, Y.-J. Lee, H. Choi, Y.D. Chung and B. Moon, "Parallel Data Processing with Mapreduce: A Survey," ACM SIGMOD Record, vol. 40, no. 4, pp. 11-20, 2012.

[10] J. Dean, and S. Ghemawat, "MapReduce: simplified data processing on large clusters", in Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation - Volume 6, OSDI'04, pp. 10–10, 2004.

[11] Dweepna Garg, Khushboo Trivedi, B. B. Panchal , "A Comparative study of Clustering Algorithms using MapReduce in Hadoop", International Journal of Engineering Research & Technology, Vol. 2.

[12] Fanyu Bu, Zhikui Chen, Peng Li, Tong Tang, and YingZhang, "A High-Order CFS Algorithm for Clustering Big Data", Hindawi Publishing Corporation Mobile Information Systems,2016

[13] Jiawei Yuan and Yifan Tian, "Practical Privacy-Preserving MapReduce Based K-means Clustering over Large-scale Dataset", IEEE Transactions on Cloud Computing, no. 99, 2017

[14] PengLi , Zhikui Chen, Laurence T.Yang, Liang Zhao , Qingchen Zhang, "A privacy-preserving high-order neuro-fuzzy c-means algorithm with Cloud computing", journal of Neuro computing, pp.1-8, 2017.

[15] Ludwig, Simone A. "MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability," International journal of machine learning and cybernetics 6, no. 6, 923-934, (2015).

Authors

M. S. Premalatha received BSc degree in Computer Science from Nesamony Memorial Christian College, Marthandam. She received Master of Computer Applications from Bishop Heber College, Thiruchirapalli and Master of Philosophy in Computer Science at Manonmanium Sundaranar University, Thirunelveli. She is currently working as Assistant Professor in the Department of Computer Applications, Nesamony Memorial Christian College, Marthandam. She is a Research Scholar in Computer Applications at Manonmanium Sundaranar University, Thirunelveli. Her field of interest is Mobile communications, Green computing and Cloud computing.

Dr. B. Ramakrishnan is currently working as Associate Professor in the Department of Computer Science and research Centre in S.T. Hindu College, Nagercoil. He received his M.Sc degree from Madurai Kamaraj University, Madurai and received Mphil (Comp. Sc.) from Alagappa University Karikudi. He earned his Doctorate degree in the field of Computer Science from Manonmaniam Sundaranar University, Tirunelveli. He has a teaching experience of 30 years. He has 23 years of research experience and published more than 70 research articles in reputed international journals (14 Science Citation Index Expanded research articles and 25 SCOPUS indexed research articles). Further, he has authored a book titled "Vehicular Ad Hoc Network and Web Vehicular Ad Hoc Network an Overview" published by the International book publisher LAP Lambert Academic Publishing with the ISBN:978-3-330-02628-5. His research interests lie in the field of Vehicular networks, mobile network and communication, Cloud computing, Green computing, Ad-hoc networks and Network security.