ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs28822019.pdf
https://doi.org/10.30534/ijiscs/2019/28822019

Cloud Computing With Big Data Clustering Using Privacy Preserving High-Order Possibilistic C-mean Algorithm

Sreelakshmy D Unni[1], Nimmymol Manuel[2]
[1]Mtech Student, India,sreelakshmyammu02@gmail.com
[2] Assistant Professor, India, nimmymol.manual@mangalam.in

ABSTRACT

In this era, Pattern Recognition and Data Mining are widely used, because the size of data is increasing day by day. The Clustering of data's are commonly done by PCM but as the size of data increases PCM is not able to do the clustering of heterogeneous data. So the Fuzzy Clustering which allows C-mean clustering can be used. The Fuzzy Clustering uses multiple Clusters data points, with each of the data points degree is varying. The Method which uses Clustering processes and image analysis is PCM for C-mean. Mainly two type of Clustering is used, one is HOPCM for clustering of large heterogeneous data and other is Normal Clustering PCM.HOPCM used for very large amount of data based on the Map Reduce functionality. Algorithm used to protect private data is PPHOPCM on Cloud by applying BGV Encryption Scheme on HOPCM data. As the experiment we can say that the heterogeneous data Clustering can be done by PPHOPCM with securing the private data using Cloud Computing. In this Paper we are giving a huge data to process to HDFS. Applying Map Reduce algorithm to the data. Then it is clustered and applying double encryption. The spark is used a tool for clustering. The encryption using BGV and AES128 Encryption Algorithms [2]. And Stored to Cloud. The encrypted data is Secure in the cloud. Using the same Encryption Key We can download the data. This improve the Security of data and Computing will be smoother.

Key words: Cloud computing, AES Encryption, HDFS, Security, Map Reduce, Cluster, BGV, AES 128, Spark.

1. INTRODUCTION

Nowadays the technology is developing day by day. The size of data to be stored in cloud will be increasing. Cloud computing makes the resource storage and the computing for large amount of data. Availability of Cloud will be direct to active users. Clustering splits the data set into meaning full clusters and sub clusters. Clustering based on similar type of data. It is a type of unsupervised classification. Clustering is done using tool Spark. It is fast and used for general purpose computing system, It is provides high level application interface java. Hadoop is used for distributed processing for big data. It can be used in clustered systems. It consist of HDFS and Map Reduce for processing. Map Reduce can be uses user defined languages. In this we are using Java. Map reduce can process large amount of data.

The data is encrypted and stored. Here we are using double encryption scheme. BGV Encryption is used first, in this encryption is homomorphic and allow computation on cipher text. The encryption algorithm we are using here is AES18 Encryption algorithm. All of the organizations have their own private data's to keep secured, then we use encryption. AES is efficient encryption algorithm and is unbreakable when compared to other algorithms. AES is symmetric key encryption with separate key is used to for both encryption and decryption [3]. The plain text chosen will be of length 64bytes to 52 bytes and 6 rounds with a complexity 2 126.8 encryptions. Key size of AES can be varied such as 128, 192, and 256. In this public key is used to perform encryption and private key used to perform the decryption.

2.RELATED WORKS

Characterises features of the big data deals HACE theorem in the perspective of data mining. The demand-driven aggregation of information sources, mining and analysis, user interest modelling, and security and privacy considerations involved by data-driven model. Diverse data sources are huge and heterogeneous. Distributed and decentralized control are Autonomous. Data and knowledge associations are complex.Relationships between samples, models, and data sources are complex. In the tensor space objective function used for high-order PCM algorithm (HOPCM) for big data clustering by optimizing. Clustering is used to separate objects into various groups according to special metrics. So that to improve the efficiency for big data clustering.Without affecting the privacy and High scalability. Not good for large data set.The fuzzy adaptive resonance theory base on investigation and three algorithms have linear computational complexity, uses a single parameter. The data clusters is identify by vigilance parameter, and are robust to modest parameter settings.Clustering mechanism is improved [7].

Cluster boundaries are better.Performance and better noise immunity are comparable.Parameter are free clustering algorithms.Initial vigilance value is zero. Probabilistic misbehavior detection scheme, for secure DTN routing toward efficient trust establishment. A periodically available trusted authority to judge the node's behavior based on the collected routing evidences and probabilistically checking on the basic idea of itrust.Detection overhead, transmission overhead decreased.

Exploit credit-based and incentive schemes to stimulate rational nodes or revocation schemes to revoke malicious nodes[1]. A high-order possibility c-means algorithm for huge amount of data set. HOPCM relation between heterogeneous data and uses map-reduce concept. A privacy-preserving HOPCM algorithm (PPHOPCM) to protect the BGV encryption scheme to HOPCM private data on cloud. The efficiency for big data clustering is increased. The privacy on cloud is preserving using privacy-preserving hopcm algorithm. The more cloud servers, making the data more suitable for clustering big data. Scalability on heterogeneous data se is reduced.

3. PROPOSED SYSTEM

Now a days many enterprise store and outsource data through cloud. By this the sensitive data's stored in cloud are much secure. The data have to keep secure is a very big problem. So that data's are encrypted and stored in cloud, data's such as financial, personal and government data's. In the encrypted stored data we do the computations. The user will request for the data and get data by giving the encryption key without much knowledge about the encrypted cloud [4].
All files are encrypted, storage and computing is not a simple job in very large systems [9]. The cost of computing have been reduced in the encrypted cloud. With the high security the document have to select with the key over encrypted cloud. [10]. Clustering improves the storage and so double encryption usage will be highly secure. [4].

In the system Authentication Flow the have to generate a ticket using TGT (Ticket Generating Ticket) to the Kerberos distribution center. When the user need type secret key and we have to run this using this system.

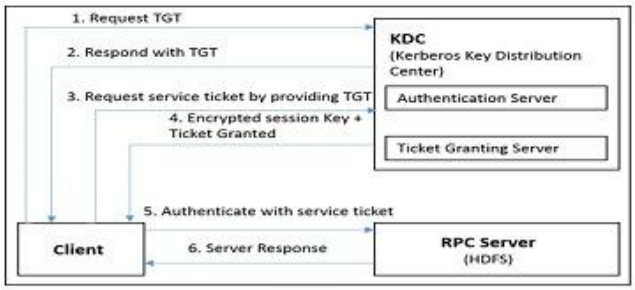

Figure 1: Systems Authentication Flow

The data file downloaded from cloud with a secret key send by system. This improves the security of the system and unauthorized access to the data files.
The encryption is done by BGV and AES128.Step by step System Authentication Flow.

1. An authorized use Request for the Ticket Generation.
2. The system gives Response with the Ticket Generated.
3. Then User request for the service that is to download a specific set of data from cloud using TGT is done.
4. The Encrypted session key and ticket is granted to the user by KDC.
5. Server authenticates the key entered by user and stored key ids same or not.
6. If the user entered key is same as the stored key the data will be decrypted and downloaded from Cloud.
7. Else the data will not be downloaded.

Map Reduce

Map Reduce is uses different languages Java, Ruby, etc. This can use large scale data clustering and analysis on different systems. It consist of two phases Map phase and Reduce Phase. Corresponding key and value pairs are there. This contain 4 phases Splitting, Mapping, Shuffling, And Reducing. Mapping is mandatory and Reducing is optional. Hadoop divides each Job into Small tasks and the input to this is those tasks

BGV

The cryptosystem supports computations on cipher texts**.** The encrypted input to produce encrypted output. So this can run on untrusted third party environments also. Mainly used for cloud data computations. Fully homomorphic encryption scheme and can use for secure data also.

AES128

AES128 is a symmetric key encryption technique which consist of 4 encryption stages. The stages are Substitution Bytes, Shift Rows, Mix Columns, and Add Round Key. AES require block size to be 128 bits, the state array of different size block has only 4 rows in other ciphers. The number of columns depend on size of the block. The data on cloud is encrypt using secret key with AES as encryption algorithm. The user have to give 16 bytes of secret key twice to uniform the secret key [1].

Table: 1: AES Versions

Version	Key Length	Block Size	No: of Rounds
AES -128	4	4	10
AES -192	6	4	12
AES -256	8	4	14

Once a file is encrypted using a secret key, the decryption have to be performed using the same secret key which have been used for encryption. If the secret key changes or misplaced then the decryption cannot be performed. In this we have to select the file for encryption. The file selected have to upload to cloud in the encrypted format. Then the file is stored in the cloud. When the user needs the file the system ask the secret key and if the secret key is same the data is decrypted and downloaded to the user. The key is given to authorized user only. With the help of a key we can ensure the security of data [8]. Only by inserting the key we can download the file from the cloud.

The 4 Stages of AES128 [5]

1. Sub Bytes transformation is a nonlinear byte substitution for each block of data.

2. Shift Rows transformation cyclically shifts the bytes within the block.

3. Mix Columns transformation groups 4bytes together forming 4-term polynomials with a fixed polynomial mode.

4. Round Key transformation adds the round key with the block of data.

Steps of Hadoop Encryption and Map Reduce Architecture

1. The input file is split and encrypted by Hadoop client.

2. The encrypted files are send to Name node.

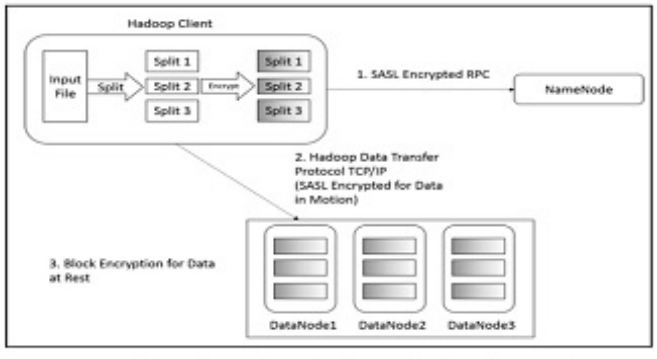

Figure 2: Hadoop Encryption and Map Reduce Architecture

3. From the Name node the data is send to data nodes for computation.

4. The data can be used by user by downloading the data with the secret key.

4. EXPERMENT AND RESULTS

The three encryption algorithms are symmetric block cipher. The AES uses 128,192,256 bits key length which can use variable length keys [8]. The DES uses 56 bit key length and triple DES uses 168 and 112 key length used by Triple DES [6].Block size of data also used by AES is 128, 192, 256 bits. DES uses 64 bit data block size. Triple DES also use 64 bit of data.There is graph Fig 3 which compares the security or usage of these three algorithms. From this we can conclude that AES is better than DES and Triple DES[5].The Encryption Algorithm BGV is also using here.The algorithm is Homomorphic and we can do computations on the encrypted cloud.So the the decryption time before computation and private data will be secured.

In this system same data is given in different types and we check the computation efficency. Also different size of data is given and we find out the variations. Fig 4 shows comparison between computation time and sze of data. As the size of data increase the data is split and computation is done[3].As size and length of data increses computation also increses.

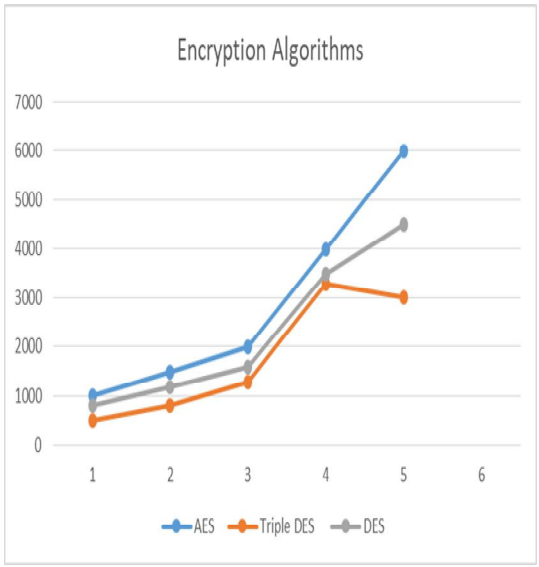

Figure 3: Comparison of encryption algorithm

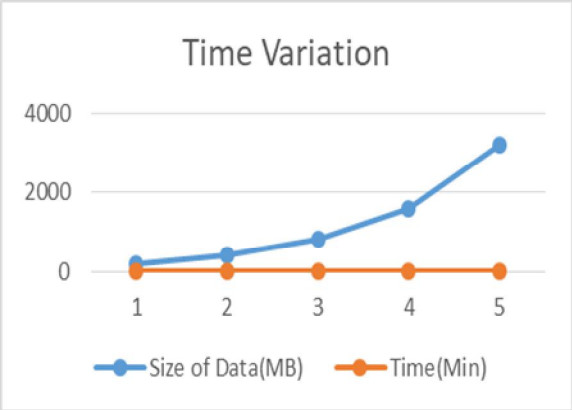

Figure 4: Computation Time Analysis

5. CONCLUSION

Nowadays the data security issues are increasing day by day in cloud. Computing data in the encrypted cloud is a complex problem. In this for encryption we are using AES and BGV encrypted cloud data. The combination of AES and BGV will have high efficiency with the data security, storage and efficiency in computing data in clustered cloud. In this HDFS and Map Reduce used for clustering of data. The performance of the system is improved as the computing time. The AES encryption is much secure than the other encryption algorithm. So that the system will be highly efficient and data is secure in the cloud data.

REFERENCES

[1]A. SCHNEIDER"WEIGHTED POSSIBILISTIC C-MEANS CLUSTERING ALGORITHMS," IN PROCEEDINGS OF THE 9TH IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS, 2000, PP. 176-180.

[2]Q. Zhang, L. T. Yang, and Z. Chen, "Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning," IEEE Transactions on Computers, vol. 65, no. 5, pp. 1351-1362, May 2016.
https://doi.org/10.1109/TC.2015.2470255

[3] R. Krishnapuram and J. M. Keller,"The Possibilistic c-Means Algorithm: Insights and Recommendations," IEEE Transactions on Fuzzy Systems, vol. 4, no. 3, pp. 385-393, Aug. 1996.https://doi.org/10.1109/91.531779

[4]S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.
https://doi.org/10.1126/science.1215842

[5]https://www.vocal.com/cryptography/advanced encryption-standard-aes.

[6]D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.
https://doi.org/10.1038/455047a

[7]L. Meng, A. Tan, and D. Xu, "Semi-Supervised Heterogeneous Fusion for Multimedia Data CoClustering," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2293-2306, Aug. 2014.
https://doi.org/10.1109/TKDE.2013.47

[8]X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data Mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.
https://doi.org/10.1109/TKDE.2013.109

[9]A. Schneider,"Weighted Possibilistic c-Means Clustering Algorithms," in Proceedings of the 9th IEEE International Conference on Fuzzy Systems, 2000, pp. 176-180

[10]M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1999, pp. 49–60.
https://doi.org/10.1145/304181.304187