International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
January-2018 Volume 5, Issue-1
Email:editor@ijermt.org www.ijermt.org

FAST DATA RETRIEVAL SYSTEM WITH SECURE DATA TRANSMISSION AND STORAGE PROCESS FOR BIG DATA IN CLOUD ENVIRONMENT

Gurpreet Singh
gurisingh0504@gmail.com
Guru Nanak Dev University, Amritsar

Abstract

Improving of technologies and increased usage of network applications lead to enhance the performance of web service system. To manage this, data analysis technique requires large storage space and high security system. Since, the technologies in the web service system lag in the data retrieving performance and secure data transmission process. To overcome these limitations, this proposed work implements data clustering and similarity based ranking system for fast retrieval of data from the bulk database in cloud environment. In this work, the performance of the proposed Eigen Score based Clustered Data Query Retrieval (ESCQR) model can be measured by the parameters such as Precision, Recall, F1-Measure and the Execution Time for retrieving and encryption of data. Here, the values ofprecision, recall and F-measure are calculated to find the accuracy level of proposed work from extracted data compare with actualdata in the database. The increasing value of these parameters represents the increase in level of matched and correctly indexed retrieved data for the given input of query request from user. Also, it defines the accuracy level of data clustering and similarity identification in the data mining process.

1. INTRODUCTION

In the era of Information Technology, a large amount of data is getting generated to the digital world. The amount of data generated is expected to double every two years, from 2500 exabytes in 2012 to 40,000 exabytes in 2020. The exponential growth of data is now leading to the beginning of a revolution that will touch every aspect of human beings' life. Big Data is about analyzing large and complex data sets to learn new and previously unknown details. The "Big Data" is often referring to the huge amounts of digital information, companies and governments collect about human beings and our environment. Big Data is a collection of any data so large and complex that it surpasses the handling capacity of traditional data management systems and techniques.

The unpredictable nature of Big Data is basically determined by the unstructured nature of a significant part of the Data that is created by present-day advances in technologies. Naturally, the pace of development of Big Data will keep on expanding for years to come. Since a large set of data is available for analysis, everybody is more focused on gaining insight into the data for making more profit and less concerned with the security problems of Big Data. In this thesis, we will characterize a novel and efficient way to deal with store and recovery of sensitive data without trading off with protection of client.

2. BIG DATA

Quick Google search brings to light millions of pages each with their own spin on the definition of Big Data. It is not a term rather a keyword big is added before data. Big Data is increasing enormously nowadays. The amount of data is exploding at an extraordinary rate as a result of developments in different Web technologies, social media, mobile and sensing devices, and satellite communications. As the user requirements/demands are increasing, more and more data is contributed towards data flood. McKinsey GI defined Big Data as large size datasets whose processing is not in the scope of typical database software tools. Knowledge discovery and better decision making are possible if information is fetched efficiently from this large collection of data. Big Data is a collection of large data sets that can't be processed using traditional tools and techniques. Due to increase in data at colossal rate a number of challenges are faced by traditional system.

Extraction of knowledge from huge volume of data has become an interesting and helpful phenomenon across the world during couples of years as different developed countries are focusing on Big Data research. Big Data is nothing new but it is an immense term refers to different data sets which are available in huge amount, heterogeneous formats including structured, unstructured, semi-structured, rapid growth with fast pace at all the time. Advancement in technology found to be responsible for the growth of data at an unprecedented rate. With the explosion in this amount of data, different new terms have also been introduced in the literature of such as "data exhaust- data added to pool of data through real life day to day activities." "Data flood", "data-lake" etc. This exponential growth of data has brought different organizations, businessmen in typical situation where decision are delayed due to inability in getting insights from this huge pool to identify the valuable information within acceptable time limits and fixed amount of resources. It is also observed that availability of more information doesn't convey more fruitful data. The huge information that needs to be fetched has some significant highlights of gigantic, high dimensional, heterogeneous, unstructured formats, noisy, outliers, inadequate, disorderly where time honoured approaches could not efficiently manage this information

due to lack of scalability, limited storage capacity, rigid data management tools expensiveness. Inability to get fitted into already existing database architecture, forces for different ways to overcome this issue.

3. PROPOSED MODEL

The work-flow of proposed ESCQR-NAG is shown in the Figure 1.

The proposed architecture performs the following functionality to achieve the better web service system for big data analysis.

1. Pre-processing using Collaborative filtering
2. Affinity Propagation for Data clustering
3. Eigen Trust Ranking
4. Similarity Estimation
5. Non-Abelian Group based data encryption

According to similarity parameters, the data retrieval is enhanced with proper paging and clustering process. The data management in web service is achieved by the communication between server and the cloud service providers with the index allocation for that clustered data and updated at every insertion of data into the database. Both the big data analysis and high data security with user details improves the efficiency of web service system. The overall process of data mining is performed by the similarity estimation of data base in the cloud environment. The detailed description of proposed algorithms with functioning steps are explained in the following descriptions.

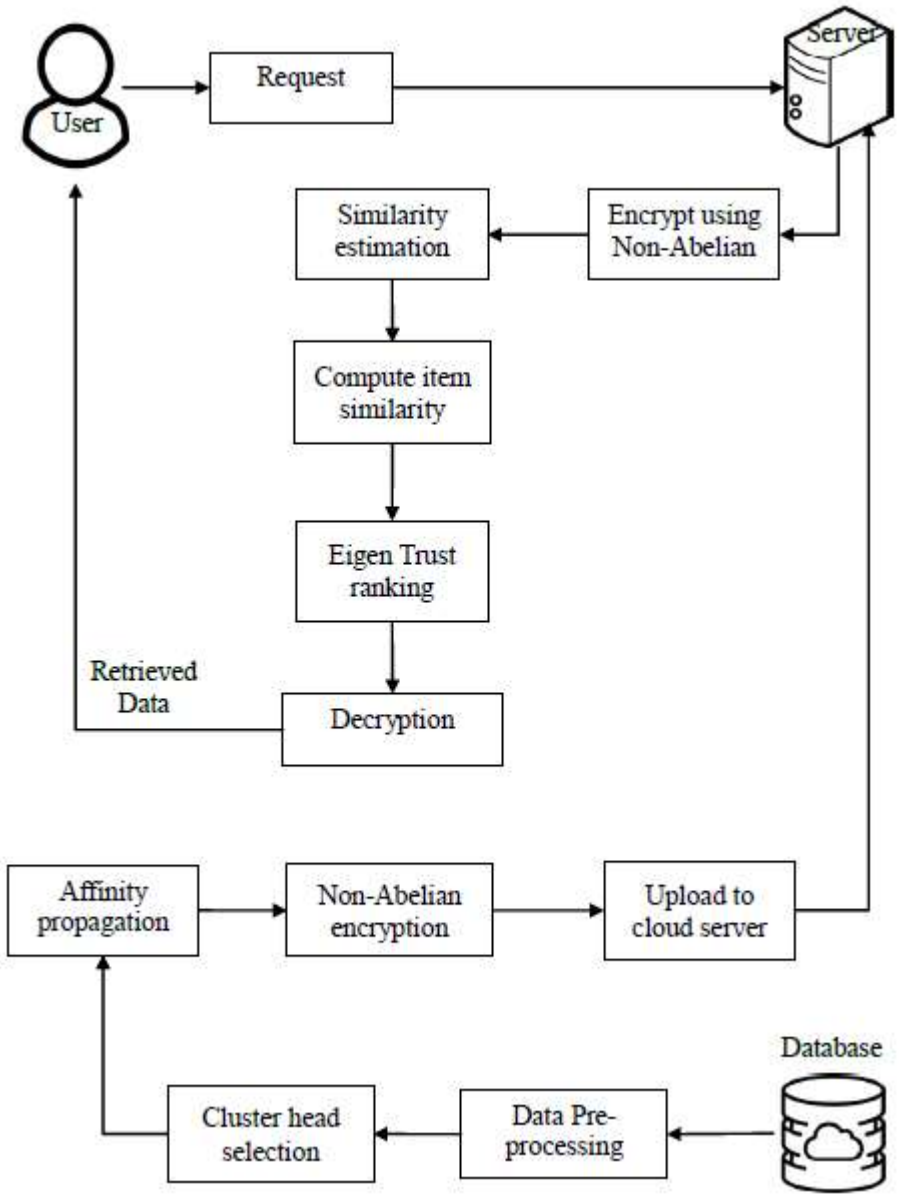

Figure 1: Work flow of proposed ESCQR-NAG

4. RESULTS & DISCUSSION

In Figure 2, it shows the precision value of proposed work in various domain of database with different data samples. Precision defines the evaluation of amount of positive prediction of retrieved data which are intersect with the actual relevancy of database and amount of mismatched positive data samples. The data samples that are used for the testing performance result in the range of 0.6 to 0.9 which represents the 60% of data to 90% of data from overall dataset. The precision value of proposed work is increased for the increased training set of the data samples. The amount of training data increased for the retrieval process results in increased accuracy value in percentage. Here the

dataset is identified as per the domain name such as Business, Entertainment, Politics, Sports, and Technologies statement from the reviewers in common social communication environment.

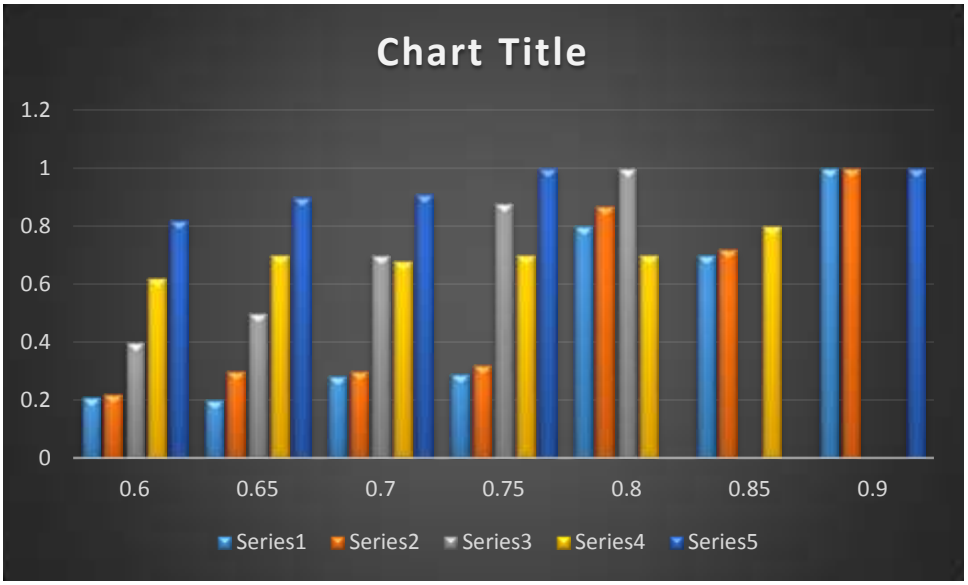

Figure 2: Precision value with different Dataset items

Then, the F1-measure is calculated by the ratio of twice the multiply of precision and recall value to the addition value of precision and recall. The F1-measure of proposed work is represented in the graph as shown in the Figure 3 in the various domains of dataset. This shows that increasing of training data samples results in increase in F1-Score that represents increase in the accuracy of proposed data clustering method.

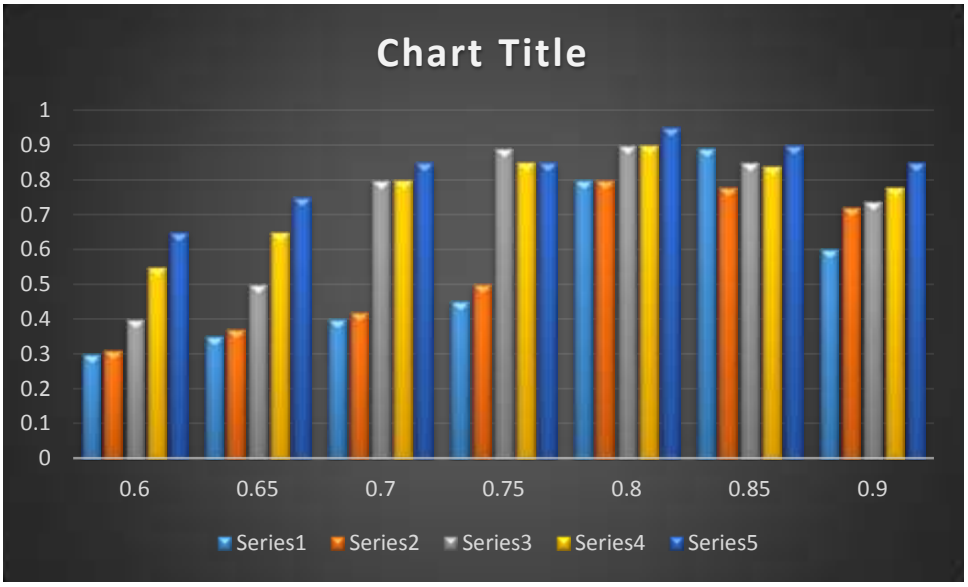

Figure 3: F-Measure value with different Dataset items

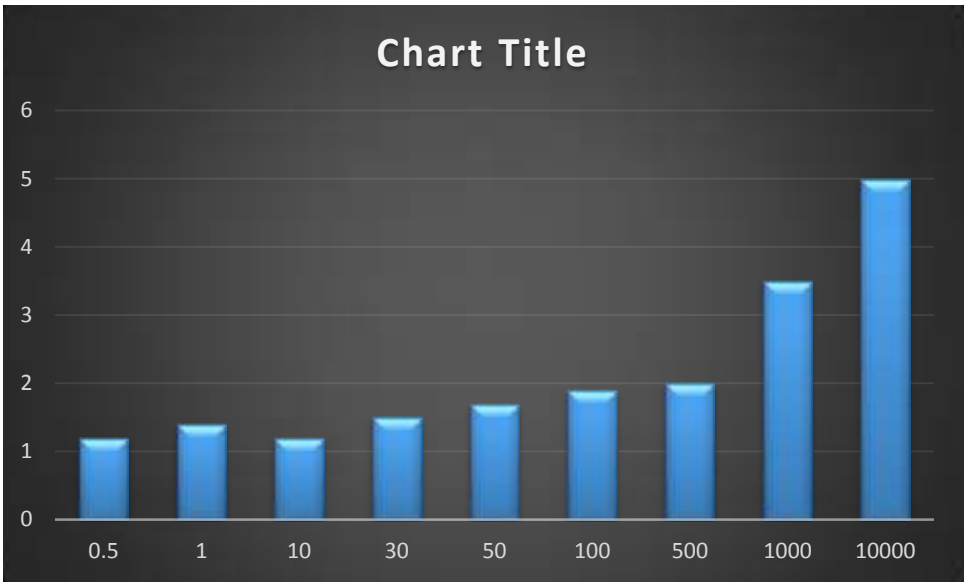

Figure 4: Encryption time analysis for different file size

The time analysis of proposed ESCQR method of data mining and the NAG method of encryption is estimated for the overall database clustering and retrieval process. In this, the time analysis is estimated for the key generation, data encryption and data decryption and key generation process of NAG algorithm.

In Figure 4, it shows the time taken to encrypt the whole dataset for different file size. In that, it consumes nearly 1 to 5 seconds for the file size in the range of 0.5 to 10000 KB. In the data retrieval stage, the decryption time is analysed for the same quantity of data while at encryption is shown in the Figure 5. In this bar plot, it shows that the decryption process consumes the time in the range of nearly 1 to 5 seconds as like the encryption stage. Increasing of file size leads to increasing the time taken for both encryption and decryption process.

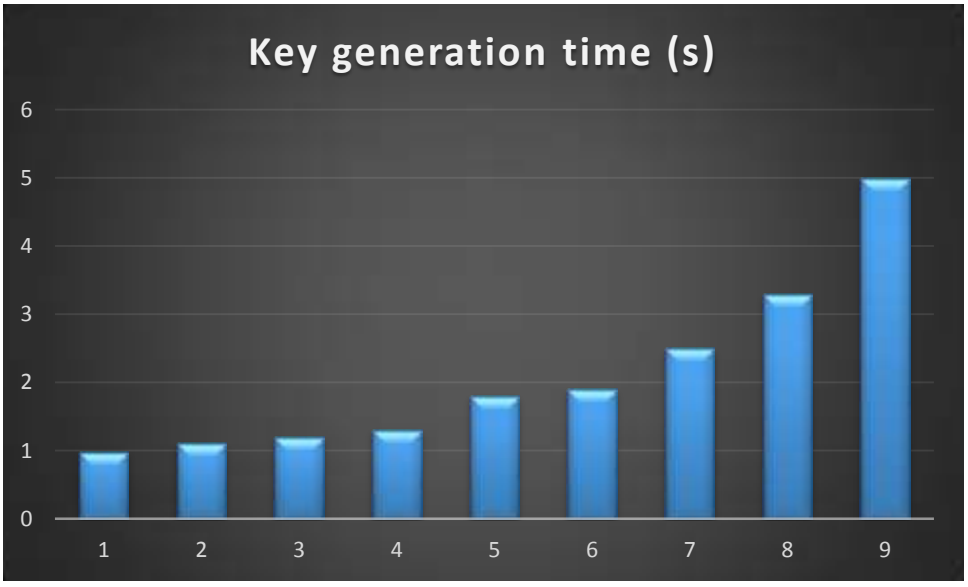

Figure 5: Decryption time analysis for different file size

Figure 6: Time analysis for different file size

In Figure 6, it shows the time analysis of key generation for NAG method for different file size. The key generation for any size of data is nearly 2 to 4 in seconds which represents that the key extraction in this NAG algorithm is normalized for all size of the data.

In these analyses, the parameter values represent the efficiency of data mining and data security process in the proposed web service system.

CONCLUSION

The recent developing technologies mainly depend on the web services of various kinds of applications that achieves huge usage among the network users. In traditional web services system, the most common issues in the data retrieval is the data clustering and match prediction between the text request from user and data stored in the database. The main aim of this paper work is to provide a fast data retrieval system with secure data transmission and storage process for big data processing in cloud environment. For the data security process, the proposed work focused in the asymmetrical approach of data encryption process. In that, ElGamal cryptosystem and Non-Abelian Group (NAG) encryption system are used to perform the encrypted data to store in the cloud environment. Since, to further enhance the secure data transmission process, the user query data and also the information about users are encrypted by those methods. During retrieval stage, the server decrypts the user requested data and their information to retrieve the required data from database. This process of preserving both request detail and the stored data results in complexity for attackers to generate the key and to extract data.

REFERENCES

[1]. Reddy, Yenumula. (2018). Big Data Processing and Access Controls in Cloud Environment. 25-33. 10.1109/BDS/HPSC/IDS18.2018.00019.

[2]. Tsuchiya, Satoshi & Sakamoto, Yoshinori & Tsuchimoto, Yuichi & Lee, Vivian. (2012). Big Data Processing in Cloud Environments. Fujitsu scientific & technical journal. 48. 159-168.

[3]. Namasudra, Suyel & Chakraborty, Rupak & Kadry, Seifedine & Manogaran, Gunasekaran & Rawal, Bharat. (2021). FAST: Fast Accessing Scheme for data Transmission in cloud computing. Peer-to-Peer Networking and Applications. 14. 10.1007/s12083-020-00959-6.

[4]. Shyla, S. & Sujatha, S.. (2022). Efficient secure data retrieval on cloud using multi-stage authentication and optimized blowfish algorithm. Journal of Ambient Intelligence and Humanized Computing. 13. 1-13. 10.1007/s12652-021-02893-8.

[5]. M. Sağir, İ. Türkeri, Bilişim Teknolojilerinin Konaklama İşletmeleriinsan Kaynaklari Uygulamalarinda Kullanimi: Konya İli Örneği. Kastamonu University Journal Of Economics & Administrative Sciences Faculty, 8, 2015.

[6]. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, Above The Clouds: A Berkeley View Of Cloud Computing, Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS, 28, 13, 2009.

[7]. P. Mell, T. Grance, The Nist Definition Of Cloud Computing, Nist Special Publication 800-1452011, 2011.

[8]. R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud Computing And Emerging It Platforms: Vision, Hype, And Reality For Delivering Computing As The 5th Utility, Future Generation Computing Systems, 25, 599-616, 2009.

[9]. R. L. Villars, C. W. Olofson, M. Eastwood, Big data: What It Is and Why You Should Care, White Paper, IDC Analyze the Future, MA, USA, 2011.

[10]. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, The rise of "Big data" on Cloud computing: Review and open research issues, Information Systems, 47, 98-115, 2015.

[11]. J. C. McCune, Data, data everywhere, Management Review, 87, 10, 10-12, 1998.

[12]. C. Tankard, Big data security, Network Security, 7, 5-8, 2012.

[13]. J. Gantz, D. Reinsel, The Digital Universe in 2020: Big data, Bigger Digital Shadows, and Biggest Growth in the Far East–United States, IDC Country Brief, IDC Analyze the Future, Framingham, MA, USA, 2013.

[14]. J. Bamford, The NSA Is Building the Country's Biggest Spy Center, Wired, 2012.

[15]. H. E. Miller, Big Data in Cloud Computing: A Taxonomy of Risks, Information Research, 18, 1, 1-19, 2013.

[16]. C. Ji, Y. Li, W. Qiu, U. Awada, K. Li, Big Data Processing in Cloud Computing Environments, 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), San Marcos, TX, USA, 17-23, 2012