International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 7, ISSN No. 2455-2143, Pages 90-97
Published Online November 2018 in IJEAST (http://www.ijeast.com)

PRIVACY PRESERVATION OF SCALABLE DATA ON CLOUD USING MAPREDUCE BASED ANONYMIZATION AND BILINGUAL SUBSTITUTION CIPHER

Sayyada Fahmeeda Sultana
Department of Computer Science
PDA College of Engineering

Dr. Shubhangi D C
Department of Computer Science
PG-Center, VTU RO, Kalaburagi, Karnataka

Abstract— **Invent of Cloud Computing technology have open scope for different organizations to use cloud for storing there huge databases like medical health details, employee personal information, voter information, etc. These databases knowingly or unknowingly may be used for data mining purposes (medical database may be mined to know who is having what disease). These databases may contain information whose leakage may harm individual identity. To solve this problem attributes in database are categorized into three groups based on their need for security as key, sensitive and quasi attributes. Key Attributes are used to identify the tuples they cannot be modified so they can be hidden, Sensitive attributes personal information need encryption, Quasi attributes may include gender, age, race, address etc. Exposure of quasi attributes can lead to identification of individuals. This paper propose a Map-reduce based Word Count approach to identify Quasi Attributes and propose to applies anonymity to achieve privacy of Quasi attributes, to provide security to sensitive attributes this paper proposes Bilingual Substitution Cipher. The proposed approach of quasi attribute identification with anonymization, and encryption of sensitive attribute through Bilingual substation cipher is fast in comparison to other state of art methods it losses very less amount of information.**

Keywords— **Bilingual Substitution Cipher, MapReduce, Word Count, Quasi attribute, Sensitive attribute, Anonymization**

I. INTRODUCTION

Cloud Computing is a highly scalable resource provider as an external service via the Internet on a pay per use basis. Cloud computing can be defined as a specialized distributed computing model, which is dynamically configured and delivered on demand [23]. This massively scalable paradigm is different from traditional networks. It is highly abstract to deliver three levels of services. The main attractiveness of cloud computing is that users only use what they need, and only pay for what they actually use. Resources are available to be accessed from the cloud at any time, and from any location through networks. There is no need to worry about how things are being maintained. These attributes describe a cloud based system as a general model providing metered on demand services to his clients. The characteristics of cloud computing are presented as follows [22]:

- On-demand self-service – cloud users may obtain resources, such as the usage of storage capacities and computing performances, without any human intervention. Similar to the principle of autonomic computing, this cloud property refers to the self managing characteristics of distributed computing resources, adapting to unpredictable changes while hiding intrinsic complexity to operators and users, in order to overcome the growing complexity of computing systems management, and to reduce the barrier that complexity raises.
- Broad network access – the large variety of heterogeneous devices, such as mobile phones, PCs, tablets, and all hand-held and static equipments have to be able to access to cloud services through standard mechanisms. This cloud ubiquitous network access characteristic is usually supported using standard protocols via Internet.
- Shared resources – based on a multi-tenant model, cloud resources are shared among several users. Note that there are no resources dedicated to a specific client. These shared capabilities are assigned, allocated and reassigned as needed to the requesting entities. The shared resources property is almost supported by several providers based on virtualization techniques, where multiple Operating Systems (OSs) co-reside on the same physical machine. However, Virtual Machines (VMs) coresidence has raised certain security requirements, namely data and process isolation.

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 7, ISSN No. 2455-2143, Pages 90-97
Published Online November 2018 in IJEAST (http://www.ijeast.com)

- Elasticity – along with self provisioning resources, cloud is characterized with the major capability to efficiently locate and release resources. This property demonstrates a scalability of greater resources. The resources are abstracted to cloud users in order to appear as unlimited and suitable. One key vulnerability that must be considered is the bandwidth under provisioning, of which malicious users can take advantage to target a service or an application availability, through Denial Of Service (DOS) attacks. Therefore, the scalability and network reliability remain important key factors to guarantee the elasticity characteristic of a cloud model.
- Metered service – this property refers to the business model adopted by cloud based services, where users pay on a consumption basis, enabling major cost reductions. By this way, the authentication and the accountability requirements have to be considered as significant needs. Moreover, the provision of metered services is supported by several monitoring tools, in order to ensure business continuity and data investigation needs.

These attributes illustrate cloud based characteristics compared to traditional computing models, supporting more efficient and scalable services. Cloud systems can be classified based on their deployment as private, public or hybrid infrastructures. Cloud storage mainly helps small and medium scale industries to reduce their investments and maintenance of storage servers. User's data sent to the cloud is stored in the public cloud environment. Data stored in the cloud storage might mingle with other user data leading to the data protection issue in cloud storage. If the confidentiality of cloud data is broken, then it will cause loss of data to the data owners. Security of cloud storage is ensured through confidentiality parameter. To ensure the confidentiality, the most common used technique is encryption.
Encryption and decryption of complete data is time consuming task to achieve high security in less time and with less resource utilization local record scheme anonymization is used. The local-recoding scheme, also known as cell generalization, groups data sets into a set of cells at the data record level and anonymizes each cell individually [1]. The text records data are classified based on their need for security attributes as[21], Table 1 gives an example to show this classification:

- Sensitive attributes
- Quasi-identifier
- Key attributes

Table 1. Example of data classification

Key Attribute	Quasi Attribute			Sensitive Attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

In this paper, we propose a model for the problem of data recoding on cloud for privacy preservation and propose a two-phase confidentiality of data through data record anonymization of Quasi attributes using Map-Reduce and Encryption of sensitive attributes using the proposed Bilingual Substitution Cipher algorithm.
The paper is organized as, Section II gives Literature survey as related work, Section III gives a detail description of Techniques used like Map-Reduce and Substitution Cipher, Section IV presents the proposed methodology, Section V gives discussion of results and analysis and Section VI is Conclusion

II. RELATED WORK

Privacy of sensitive information related to individuals, such as medical data, personal details, employ details, etc, are collected, stored and processed in a large variety of application domains. Such data is typically used to provide better quality services to individuals and its availability is crucial in many contexts. In the case of healthcare, for example, the availability of such data helps prevent medical errors and enhance patient care. Privacy of sensitive data may have very important, data stored on cloud intentionally or unintentionally used for data mining leading to individual user privacy violation, in current state of art many techniques are used to protect the privacy based on specific domain. Anonymization is used for protecting data privacy different scenarios like [3] , k-anonymity [4], l-diversity [5], t-closeness [6], privacy preserving properties of random data perturbation techniques[7]. A genetic TDS and BUG with pseudo-identifier for privacy preservation over incremental data sets [8].
Privacy preservation of sensitive attribute is essential needs to be protected, sensitive information tends to be overly specific, thus of less utility, to classification; even though masking sensitive information eliminates some useful structures, authors of [9] proposed a top-down approach to iteratively specialize the data from a general state into a special state, guided by maximizing the information utility and minimizing the privacy specificity. The top down approach serves a natural and efficient structure for handling categorical and

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 7, ISSN No. 2455-2143, Pages 90-97
Published Online November 2018 in IJEAST (http://www.ijeast.com)

continuous attributes and multiple anonymity requirements [9]. Providing anonymization and reducing information loss in incremental dataset through grouping and local recoding, identifying quasi attribute is based on the maximum count of unique value [10] using cut-vertex [11]. P. Shyja Rose, J. Visumathi and H. Haripriya provide on comprehensive overview of privacy among the data which is placed on the public cloud [12].

Tong Yi and Min yong Shi proposed a Privacy Protection Method for Multiple Sensitive Attributes Based on Strong Rule, they took relationship between different sensitive attributes into account, presents a mixed data publishing model based on rating [13]. A rule based approach for determining the attribute's sensitivity level. Primary keys, indices and statistics on the database stored in the DBMS for optimization purpose are used to detect attributes that are quite identifying for the tuples [14].

III. TECNOLOGY USED

A. MapReduce

The MapReduce programming model is conceptually simple, based on two steps applying a mapping process and then reducing (condensing/ Collecting) the results- it can be applied to many real world problems. MapReduce is fast through the use of parallel data flow. The parallel execution of MapReduce requires other steps in addition to the mapper and reduce processes. The basic steps are as follows [24]:

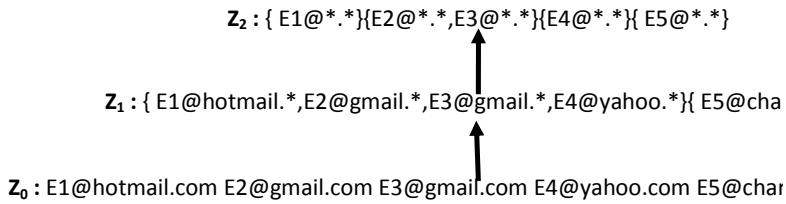
- Input Splits: The input splits used by MapReduce are the logical boundaries based on input data. The number of splits corresponds to the number of mapping processes used in the map stage.
- Map Step: The mapping process is where the parallel nature comes into play. For large amounts of data, many mappers can be operating at the same time. The users provide the specific mapping process.
- Combiner Step: It is possible to provide an optimization or pre-reduction as part of the map stage where key-value pairs are combined prior to the next stage. The combiner stage is optional.
- Shuffle Step: Before the parallel reduction stage can complete, all similar keys must be combined and connected by the same reducer process. Therefore, results of the map stage must be collected by key-value pairs and shuffled to the same reducer process.
- Reduce Step: The final step is the actual reduction. In this stage, the data reduction is performed as per the programmers design.

B. Data Anonymization

Data anonymization seeks to protect private or sensitive data by deleting or encrypting personally identifiable information from a database. Data anonymization is done for the purpose of protecting an individual's or company's private activities while maintaining the integrity of the data gathered and shared. Data anonymization is also known as data obfuscation, data masking, or data de-identification. Generalization of attributes replaces with less specific, but semantically consistent values figure 1 show the generalization of zip code, age, gender and email id.

Zip Age Gender

(a)



email ID

(b)

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 7, ISSN No. 2455-2143, Pages 90-97
Published Online November 2018 in IJEAST (http://www.ijeast.com)

Figure 1. Anonymization of attributes (a) zipcode, age, gender (b) Email ID

C. Substitution Cipher

In cryptography, a substitution cipher is a method of encrypting by which units of plaintext are replaced with cipher text, according to a fixed system; the "units" may be single letters (the most common), pairs of letters, triplets of letters, mixtures of the above, and so forth. The receiver deciphers the text by performing the inverse substitution [2].

Substitution ciphers can be compared with transposition ciphers. In a transposition cipher, the units of the plaintext are rearranged in a different and usually quite complex order, but the units themselves are left unchanged. By contrast, in a substitution cipher, the units of the plaintext are retained in the same sequence in the cipher text, but the units themselves are altered.

There are a number of different types of substitution cipher. If the cipher operates on single letters, it is termed a simple substitution cipher; a cipher that operates on larger groups of letters is termed poly-graphic. A mono-alphabetic cipher uses fixed substitution over the entire message, whereas a poly-alphabetic cipher uses a number of substitutions at different positions in the message, where a unit from the plaintext is mapped to one of several possibilities in the cipher text and vice versa.

Example of substitution cipher is shown in figure 2, cipher alphabet by shifted alphabets (creating the Caesar and Atbash ciphers, respectively) or scrambled in a more complex fashion, in which case it is called a mixed alphabet or deranged alphabet.

Original	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Ciphered	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M

Figure 2. Substitution Cipher

IV. Proposed Methodology

The proposed methodology is based on two phases

- Privacy Preservation of Sensitive Attributes
- Privacy Preservation of Quasi Attributes

A. Privacy Preservation of Sensitive Attributes

DEFINITION 1. *Sensitive attributes: Databases have set of attributes Sensitive attributes* $S = \{s1, \dots, sk\} \subseteq D$ *are attributes that are used to evaluate the risk derived from exposing the data. The sensitive attributes are mutually excluded from the quasi-identifier attributes (i.e.,* $\forall S \cap Q = \emptyset$ *where Q is a set of quasi attributes).*

Sensitive attributes are the attributes whose exposure to non authorized people may cause problem to the person whose identity is know through key attribute. To provide privacy to Sensitive attributes we proposed a Bilingual Substitution Cipher algorithm. Bilingual Substitution Cipher algorithm is a combination of two languages used alternately to encrypt text data. Algorithm 1 gives the Bilingual Substitution Cipher algorithm.

Algorithm 1 Bilingual Substitution Cipher

Input: Selected Sensitive attributes value from dataset
Output: Encrypted Sensitive attribute value
Step 1: Create two tables G and F of characters with different pronunciations for same characters [ex. A(in English)=EIN (in German), UNE (in French), B(in English) = BEN(in German), BEH(in French) , …]
Step 2: Find the index of sensitive attribute
Step 3: Repeat for each v_i *value of sensitive attribute* $i = 1,2, \dots n, n = number\ of\ rows\ in\ dataset$
Repeat for each character $c_j\ in\ v_i, where\ j\ is\ word\ length$
if ASCII value of c_j *odd*
replace c_j *with "\" and proceed by corresponding character in table G to get* c_j'
else
replace c_j *with ":" and proceed by corresponding character in table T to get* c_j'
end

B. Privacy preservation of Quasi Attributes

DEFINITION 2. Quasi attributes. In a database $D = \{A1, \dots, An\}$ *the Quasi attributes is defined as* $Q = \{qi1, \dots, qik\} \subseteq \{Ai, \dots, An\}$ *are attributes that can be linked, possibly using an external data source, to reveal a specific entity that the specific information is about.*

Privacy preservation of Quasi Attributes by first identifying quasi attribute using algorithm 2, algorithm 2 is based on MapReduce Word Count program to identify quasi attributes

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 7, ISSN No. 2455-2143, Pages 90-97
Published Online November 2018 in IJEAST (http://www.ijeast.com)

the quasi attributes are identified by seeing the key value pair output of word count program. K indicates the number of values a quasi attributes take. The paper propose a systematic approach to find value a k, as the choice of the k value in the anonymity problem can be made in a much informed manner rather than arbitrarily [15]. Figure 3 show the work flow of MapReduce Word Count to identify quasi attribute is performed by generalization of quasi attributes using algorithm 3 Anonymization of quasi attributes in this algorithm 3 quasi attributes are categorized into numerical and character set attributes, numerical attributes are anonymized using roundoff strategy and Character set attributes are anonymized using algorithm 1 Bilingual substitution cipher. Algorithm 1 Bilingual Substitution cipher uses the idea of alphabets with different pronunciation.

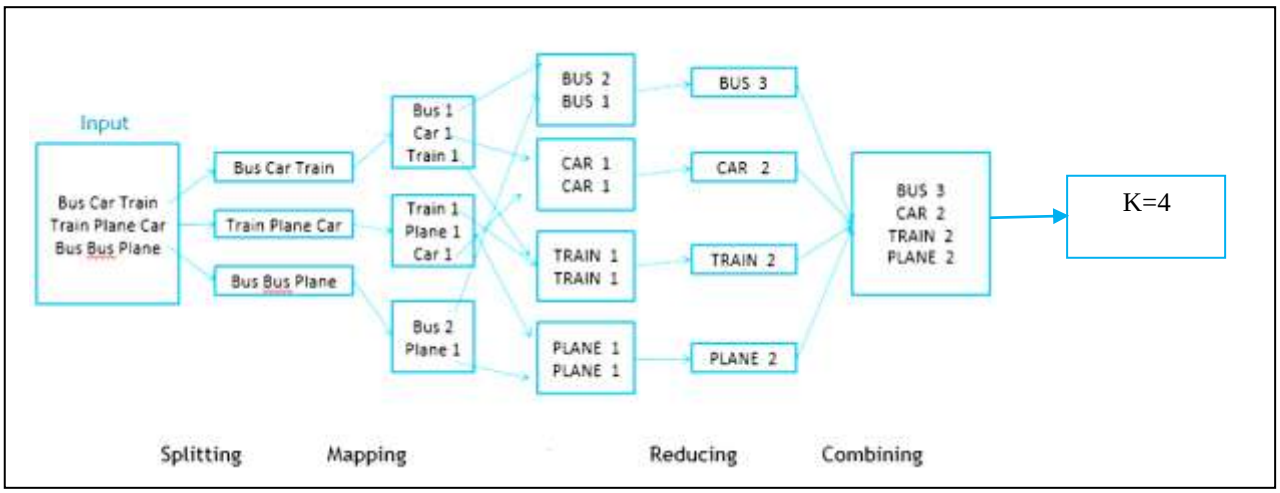


Figure 3 Work Flow of MapReduce Word Count to identify Quasi attribute

Algorithm 2. Identification of Quasi Attributes
Input: Scalable Text Data(or Database) D={A_1,A_2,.... A_n}, Where A_i is i^{th} attribute
Output: Identified set of Quasi Attributes D_Q, Set of Key attributes D_s
Step 1: For each A_i attribute run MapReduce Word Count job to get K for A_i
Step 1.1: if K is equal number of records in D, then A_i is key attribute D_k={ A_i}
Else if K is less then number of record in D, then A_i is quasi attribute D_Q={ A_i}
Step 1.2: repeat Step1 for each attribute in D
Step 2: Stop

Algorithm 1 IQA is executed on each attribute if an attribute is key attribute then it maintains uniqueness property hence the count K from word count job is equal to number of records in database D. For Quasi and Sensitive attributes value of K is less than number of records in relation.

Algorithm3. Anonymization of Quasi Attributes (AQA Algorithm)
Input: Quasi Attributes D_Q={A_1, A_2, … , A_m} where m is number of Quasi attributes
Output: Anonymized Quasi Attributes D_Q^1={ A_1^1, A_2^1, … , A_m^1}, Set of Sensitive attributes D_s

Step 1: Categorize Quasi attributes based on its value as set of Numerical Quasi attributes NA and set of character attributes CA
Step 2: For each Numerical Quasi Attributes NA_i where i=1,2,.. p, p is number of numerical quasi attributes
Step 2.1: Choose the level of anonymization as m
Switch (m)
{
Case 1: Step a. for each tuple value v of NA_i
Step b. Find the digit at one's position as o in v
if o>5 replace o in v by 9 else replace zero
break
Case 2: Step a. for each tuple value v of NA_i
Step b. roundoff(v) to nearest 10
break
Case 3: Step a. for each tuple value v of NA_i
Step b. roundoff(v) to nearest 100
break
Case 4: Step a. for each tuple value v of NA_i
Step b. roundoff(v) to nearest 1000
break
. . .
}

Step 3: For each Character Quasi Attribute CA_i, where i=1,2,…q, q is number of character quasi attribute
Step 3.1: Add CA_i to D_s

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 7, ISSN No. 2455-2143, Pages 90-97
Published Online November 2018 in IJEAST (http://www.ijeast.com)

Step 3.2: Run Bilingual Substitution Cipher on D_s
Step 4: Replace each D_Q and D_Q^1 anonymized

The proposed Algorithm 3 AQA works for anonymizing numerical quasi attribute by roundoff to nearest tens, hundreds, thousands and soon, depending upon the number of digits in value t. AQA Algorithm 3 also generate D_s, which

are the character based attribute assigned as Sensitive attribute, we propose a new substitution cipher algorithm Bilingual Substitution Cipher algorithm

V. EXPERIMENTAL EVALUATION

In order to validate our proposed methodology, three datasets are used for experimental analysis. These datasets are publicly available we used the training set of the UCI Adult dataset [17], and samples of the publicly-available voter lists from Florida [16] and Michigan [18] datasets. Table 2 lists the datasets and their respective attributes as used in experiments here. Table 3 shows the separated attributes based on their need for security as key attributes, quasi attributes and sensitive attributes. Quasi attributes are identified using MapReduce based Word Count program as illustrated in algorithm3.

Table 2. Dataset Attributes

	Adult Dataset	**Florida Dataset**	**Michigan Dataset**	**Random logic dataset VBA**
Number of records:	**32 440**	**134 791**	**72 825**	**1000**
Attributes	**Patient Id, Name ,Gender, Race, Marital Status, Height, E-mail, Year of Birth**	**Name, Surname ,ZIP code, Gender, Year of Birth, Race, Height, Address, E-mail, Phone Number, Year of Birth**	**Name, Surname, ZIP Code, Gender, Year of Birth, Weight, Address, Year of Birth**	**Emp ID, Name, Gender, E Mail, Date of Birth, Weight in Kgs., Date of Joining, Salary, Phone No. , Zip, User Name, Password**

Table 3 Identification Quasi attributes using MapReduce WordCount, separation of key and Sensitive attributes

	Adult Dataset	**Florida Dataset**	**Michigan Dataset**	**Random logic dataset VBA**
Key Attribute	**Patient Id, Name**	**Name, Surname**	**Name, Surname**	**Emp ID, Name**
Quasi Attributes	**Gender, Race, Marital Status, Height**	**ZIP code, Gender, Race, Height**	**ZIP Code, Gender, Weight**	**Gender, Weight in Kgs., Date of Joining, Zip**
Sensitive Attributess	**E-mail, Year of Birth**	**Address, E-mail, Phone Number, Year of Birth**	**Address, Year of Birth**	**E Mail, Date of Birth, Salary, Phone No. , User Name, Password**

Figure 4 (a) Selected attributes from Random Logic Dataset in original form where EMP ID is a key attribute, Prefix, Age, and Zip are identified as Quasi attribute, Password is a sensitive attribute, Figure 4(b) shows the anonymized form of Quasi data and encrypted form of Sensitive data

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 7, ISSN No. 2455-2143, Pages 90-97
Published Online November 2018 in IJEAST (http://www.ijeast.com)

EMP ID	Prefix	Age	Zip	Password
677509	Drs.	34	83445	DCa}.T}X:v?NP
940761	Ms.	46	73492	TCo\j#Zg;SQ~o
428945	Dr.	52	47252	GO4$J8MiEh[A
408351	Drs.	43	18572	0gGRtp1HfL<r5
193819	Mr.	36	57623	Rd<Y8cp!@R;*%F
499687	Mr.	22	57623	K7&5aY/*
539712	Ms.	59	37464	xJdKlAcYQhT_BE#
380086	Mrs.	43	47239	Uc+VG%vuZU<k
477616	Hon.	23	73422	K}^USc0l7[A
162402	Hon.	48	73421	3o8>v&tYxjyEAo

(a)

(b)

Figure 4. (a) Selected attributes from Random Logic Dataset (b)

EMP ID	Prefix	Age	Zip	Password
677509	2	30	83450	DfsCdfg'ah}.Tdg}X:'dvay?NdfP
940761	5	50	73500	TfgCg'o\'jfgogt:#Z'ge:;SdgQ~'go
428945	1	50	47250	GOgd4$Jdg8Mg'egeE'ash[A
408351	2	40	18570	0'gse:GR'tadfgy'pe:1gH'gfeffL<'er:5
193819	3	40	57620	R'sfdafsy<Y8'sfsay'pe:!@R;*%sfF
499687	3	20	57620	Ksf7&5'asfhfsffY/*
539712	5	60	374600	'ee-grefskJ'dsfayK'sfelA'sayYQ'ashT_BE#
380086	4	40	47240	U'say+VG%'vay'sfdooZU<'ksfa
477616	6	20	73420	K}^UarS'trgfsay0'el7[A
162402	6	50	73420	3'o8>'vay&'tadbhdyY'ee-grek'jgrosfsafgat:'dazesgdEgfsA'o

Anonymized and Encrypted form of dataset

Experiment result demonstrates that our approach of anonymization of quasi attributes and encryption of sensitive attribute though the proposed bilingual substitution cipher algorithm is able to retrieve 90% of original information through decryption and de-anonymization by the use of Microtable. encryption of proposed scheme significantly improves the capability of defending privacy disclosure risk and improves the scalability over existing approaches. Figure 5 shows the dataset size on x-axis and execution time as y-axis the proposed approach Anonymity with word count is compared with DHC (Divisive Hierarchical Clustering based MapReduce [19] and k-member clustering approach (kMC) proposed in [20] the proposed approach of anonymity with MapReduce word count achieve high speed. The experiment was executed on Amazon web services with one master and two core instances each was '4 vcore 8GIB memory', Hadoop distribution used is Amazon 2.8.5.

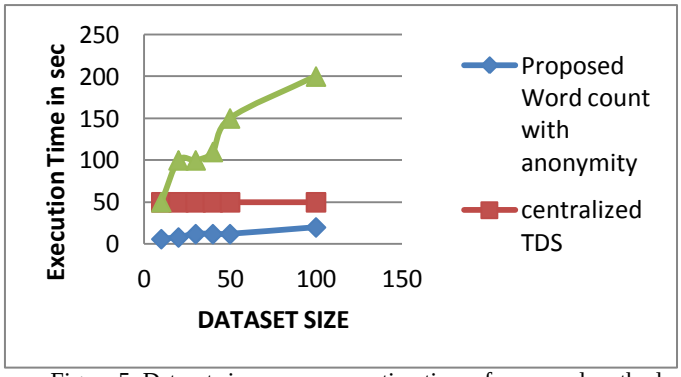


Figure 5. Dataset size verses execution time of proposed method with Centralized TDS and DHC approach

VI. CONCLUSION

The paper proposes an anonymization and encryption scheme for huge data stored on cloud that need privacy preservation. To achieve privacy of information when the data is on cloud from the perspectives of capability of defending privacy breaches, scalability and time-efficiency. The proposed privacy preserving approach for quasi and sensitive attributes of database allowing multiple sensitive attributes to undergo Bilingual Substitution Cipher encryption algorithm for encryption. The proposed a two phase Anonymization approach for quasi attributes for numerical and character attribute values. The identification of quasi attributes is performed by using MapReduce Word Count Program. The approach based on Map Reduce to address the above problem is time-efficient, with less data loss on anonymization as we anonymize character attribute values using proposed Bilingual substitution cipher, and numerical attributes can be recovery at the time of de-anonymization using micro-table. The proposed approach can be extended to provide privacy to key attributes without loss of uniqueness property of key attribute.

VII. REFERENCE

[1] Xuyun Z, Wanchun D, et.al, (2015), Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud, IEEE transactions on computers, vol. 64, issue 8(pp. 2293-2307)

[2] Bradley H, Ryan H, Grzegorz K. (2014), Solving Substitution Ciphers with Combined Language Models, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2314–2325, Dublin, Ireland (pp. 2314-2325)

[3] Ali I., Murat K., Elisa B. (2009), Using Anonymized Data for Classification, IEEE 25th International

Conference on Data Engineering. DOI: 10.1109/ICDE.2009.19

[4] Latanya S. (2002), Achieving k-Anonymity Privacy Protection Using Generalization And Suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol 10, issue 5, 2002 (pp. 571-588)

[5] Ashwin M., Johannes G., Daniel K. (2006), L-diversity: privacy beyond k-anonymity, 22nd International Conference on Data Engineering (ICDE'06), DOI: 10.1109/ICDE.2006.1

[6] Ninghui Li, Tiancheng Li, Suresh V. (2007), t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, IEEE 23rd International Conference on Data Engineering, DOI: 10.1109/ICDE.2007.367856

[7] Hillol K.,Souptik D., Qi Wang, Krishnamoorthy S. (2003), On the Privacy Preserving Properties of Random Data Perturbation Techniques, Third IEEE International Conference on Data Mining, DOI: 10.1109/ICDM.2003.1250908

[8] Sreedhar, Faruk,Venkateswarlu. (2017), A genetic TDS and BUG with pseudo-identifier for privacy preservation over incremental data sets, Journal of Intelligent & Fuzzy Systems, vol. 32 issue 4(pp. 2863-2873) DOI: 10.3233/JIFS-169229

[9] Benjamin C. M. Fung , Ke Wang, Philip S. Yu (2005), Top-Down Specialization for Information and Privacy Preservation, 21st International Conference on Data Engineering (ICDE'05), DOI: 10.1109/ICDE.2005.143

[10] R. Mahesh, T. Meyyappan (2017), Anonymization and Reducing Information loss in Incremental Dataset through Grouping and Local Recoding, Australian Journal of Basic and Applied Sciences, Vol. 11, issue 11, (pp. 139-147)

[11] YanY, WanjunW, Xiaohong H., Lianxiu Z. (2018), Finding Quasi-identifiers for K-Anonymity Model by the Set of Cut-vertex, Engineering Letters Vol. 26 issue 1 (pp.150-160.)

[12] P. Shyja Rose, J. Visumathi, H. Haripriya (2016), Research Paper on Privacy Preservation by Data Anonymization in Public Cloud for Hospital Management on Big data, I J C T A, Vol. 9 issue 7, (pp. 3095-3102)

[13] Tong Yi, Minyong Shi (2015), Privacy Protection Method for Multiple Sensitive Attributes Based on Strong Rule, Mathematical Problems in Engineering, Vol. 2015. http://dx.doi.org/10.1155/2015/464731

[14] Cedric du Mouza, Elisabeth M., et.al., (2010), Towards an Automatic Detection of Sensitive Information in a Database, 2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications, DOI: 10.1109/DBKDA.2010.17

[15] Rinku D, Indrajit R, Indrakshi R, Darrell W., (2008), On the Optimal Selection of k in the k–Anonymity Problem, 2008 IEEE 24th International Conference on Data Engineering, DOI: 10.1109/ICDE.2008.4497557

[16] Registered voters in the state of Florida, http://flvoters.com/, 2018.

[17] M. Lichman, UCI machine learning repository, 2018 Available: http://archive.ics.uci.edu/ml

[18] Registered voters in the state of Michigan, http://michiganvoters.info/, 2018.

[19] Selvi.U, Selvaprabu.D, (2016), Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud, International Journal of Computer Science and Engineering Communications, Vol.4, issue.2, (pp.1301-1306)

[20] X. Xiao and Y. Tao, (2006) ,Personalized privacy preservation, Proc. ACM SIGMOD Int. Conf. Manage. Data, (pp. 229–240).

[21] R. Praveena, M. L. Valarmathi, S. S. (2015), Attribute Segregation based on Feature Ranking Framework for Privacy Preserving Data Mining, Indian Journal of Science and Technology, Vol 8 issue 17, DOI: 10.17485/ijst/2015/v8i17/77584

[22] Thomas Erl, Cloud Computing: Concepts, Technology & Architecture, Prentice Hall

[23] Michael Kavis, Architecting the Cloud: Design Decisions for Cloud Computing Service Models (SaaS, PaaS, and IaaS), Michael Kavis O'Realy Media

[24] Srinath Perera, Thilina Gunarathne, Hadoop MapReduce Cookbook, PACKT publication, February 2013, http://barbie.uta.edu/~jli/Resources/MapReduce&Hadoop/Hadoop%20MapReduce%20Cookbook.pdf