ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

High Dimensional Data Attribute Security and Utility Using Hadoop

Shital Suryawanshi, Prof. V.S.Wadne

PG Student, Department of Computer Engineering, Savitribai Phule Pune University JSPM's Imperial College of Engineering & Research, Wagholi, Pune,India.

Assistant Professor, Department of Computer Engineering, Savitribai Phule Pune University, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India

ABSTRACT: Big data is large volume, heterogeneous, decentralized distributed data with different dimensions. In Big data applications data collection has grown continuously due to this it becomes expensive to manage, capture or extract and process data using existing software tools. Performing data analysis is becoming expensive with increasing size of data in data warehouse. Data privacy is one of the challenge in data mining with big data. To preserving the privacy of the user we need to use some method so that data privacy is preserve and at the same time increase the data utility. In existing centralized algorithms it assumes that the all data should be at centralized location for anonymization which is not possible for large scale dataset. And there were distributed algorithms which mainly focus on privacy preservation of large dataset rather than the scalability issue. In the proposed system we focus to maintain the privacy for distributed data, and also overcome the problems of M-privacy and secrecy approach with new anonymization and slicing technique. Our main goal is to publish an anonymized view of integrated data, which will be immune to attacks. We use MR-Cube approach which addresses the challenges of large scale cube computation with holistic measure. Using MR-Cube approach the resultant data is in materialized view. We can easily anonymize the materialized data for the user. The data is anonymized using slicing. Slicing contains tuple partition, generalization, slicing and anonymization. Once slicing is done the anonymized data can freely access by user.

KEYWORDS: MR-Cube, MapReduce, Data anonymization, data privacy, slicing

I. INTRODUCTION

Big data where the information comes from multiple, heterogeneous, autonomous sources with complex relationship and it is continuously growing. Up to 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [1]. This shows that it is very difficult for big data applications to manage process and retrieve data from large volume of data. It's become challenge to mine knowledgeable information from large dataset for future use. There are different challenges of Data mining with Big Data. One of them is data privacy challenge, which can be solved using different approaches like key based encryption [3] and anonymization. To process and compute high dimensional distributed data is also one of the challenge. In this the data has different dimensions for e.g. in hospitals, patients data is stored in text and images , videos are used to stored results of X-ray, CT scan for detail examinations. MR-Cube approach is used for efficient computation of cube [8].

Performing data analysis on such big data becomes expensive as the data is distributed, continuously keeps growing and the nature of data is structured, unstructured and semi structured. Pdf, videos, images are example of unstructured data. For analyzing multidimensional data, data cube is powerful tool. Consider a data warehouse maintain the sales information containing <city, country, state, day, month, year and sales>. Where city, country and state attribute are of local dimension and attribute day, month, and year are of temporal dimension. Cube analysis provides convenient way to discover insight from the data by computing aggregate measures. Top-Down approach, Bottom-Up Computation (BUC), A Mining Cubing Approach, Parallel approach is some of the cube computation techniques. There are two main limitations in the existing cube computation techniques. First they are design for single machine or cluster with

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

small number of nodes. It is difficult for businesses or companies containing huge data storage. Second limitation is many technique use algebraic measure to avoid processing groups with a large number of tuples. There is need of technique to compute large cube efficiently in parallel. MapReduce programming paradigm is used to analyze and process such large scale data.

Existing anonymization algorithms such as centralized algorithms [9] assume that all data processed should fit in memory which is not possible for large dataset. There are also some distributed algorithms which mainly aim at security integrating and anonymizing multiple data sources rather than scalability issue of TDS anonymization. In the proposed system we focus to maintain the privacy for distributed data, and also overcome the problems of M-privacy and secrecy approach with new anonymization and slicing technique.

As there is increasing need of sharing personal information from distributed database, the special care should be taken to protect it from attacker. Attacker can be single entity or group of entities. Attacker can breach privacy with the use of background knowledge. Collaborative data publishing considered as a multi-party computation problem. In this problem multiple providers want to compute an anonymized view of their data without disclosing any private and sensitive information. A data recipient that might be an attacker, e.g., q0, attempts to gather additional information about data records using the published data, D*, and background knowledge, BK. For example, k-anonymity [10] protects against identity disclosure attacks by requiring each quasi identifier equivalence group (QI group) to contain at least k records. L-Diversity requires that each QI group should contain at least l "well-represented" sensitive values. Differential privacy guarantees that the presence of a record cannot be inferred from a statistical data release with little assumptions on an attacker's background knowledge.

We considered a potential attack on collaborative data publishing. We used slicing algorithm for anonymization and L diversity and verify it for security and privacy by using binary algorithm of data privacy. For high dimensional data, slicing algorithm is useful. It divides the data in both vertical and horizontal manner. Encryption can increase security, but the limitation is there could be loss of data utility. Our main goal is to publish an anonymized view of integrated data, which will be immune to attacks. We improve the security and privacy with the help of slicing technique which fulfils privacy verification with better performance than provider aware (base algorithm) and encryption algorithm.

II. RELATED WORK

Big data is generated and collected from various autonomous, heterogeneous sources and having different dimensions [1]. The Big data are continuously growing is become the challenge to processing large data and securing that data. Ingale et a1. [3] Proposed Advance Encryption Standard (AES) with k-anonymization for privacy conserving to achieve privacy. K-anonymization allows database to maintain a suppressed and generalized form of data. A different anonymization algorithm with different operations has been proposed [9] [11]. As the dataset becomes very large that anonymizing such data becomes challenge for traditional anonymization techniques. To analyze this large amount of data various cube computation techniques [7] have been used. The existing cube computation techniques have several limitations that they compute cube over limited number of node and many techniques compute with algebraic measure only. Nandi et al. [8] proposed MR-Cube approach for efficient cube computation with holistic measure for large dataset.

Fung extended the k-anonymization algorithm to preserve the information for cluster analysis. The major challenge in this is the lack of class labels that could be used to guide the anonymization process. The solution is to first partition the original data into clusters on the original data then problem is converted into counterpart problem for classification analysis where class label encode the cluster information in the data and then apply TDS to preserve k-anonymity.
FUNG et al. [9] proposed a new privacy model LKC-privacy to overcome the challenges of traditional anonymization methods using centralized and distributed anonymization algorithm. A data structured TIPS (Taxonomy Indexed partitionS) is exploited in centralized algorithm to improve efficiency of TDS. molloy at el. [2] used slicing, which partitions the data both horizontally and vertically. Slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. Slicing can be used for attribute disclosure protection. Generalization and bucketization are

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

anonymization techniques. For generalization is used for k-anonymity and l-diversity is used for bucketization. In both of these approaches the attributes are partition into three types. The first is identifier like ID No or SSN, second is Quasi-identifier which is combination of more than one attribute and the third is sensitive attribute. For anonymous data these identifiers are first remove form data and then partition into bucket.

Identity disclosure, Attribute disclosure and membership disclosure are threat which needs to overcome. Slicing is used on high dimensional data prevent membership disclosure reduce the information loss and increased data utility. Top down specialization approach [16] used to anonymized large scale data set on cloud. Existing TDS approaches for large scale data sets having scalability problem. The centralized TDS approaches use TIPS to improve the scalability and efficiency by indexing anonymous data records. Centralized approaches suffer from low scalability and efficiency when it handles large scale data sets. The assumption in centralized approach to fit all data in memory for processing which is not possible for large data sets. To handle scalability issue we used MR-Cube approach which first generate annotated lattice and then used it to perform main MR-Cube MapReduce.

III. PROPOSED SYSTEM

a. Problem Definition:

While maintaining the data privacy in big data or in a multi-provider environment various challenges are faced, like data security, data utility and data processing. Existing provider aware algorithm was not allowing enough data utility while securing data. In the proposed system we focus to maintain the privacy for distributed data, without loss of data and also overcome the problem of M-privacy and a secrecy approach is used with new slicing technique.

Our main goal is to publish an anonymized view of integrated data, which will be immune to attacks. We improve the security and privacy with the help of slicing technique which fulfils privacy verification with better performance than provider aware (base algorithm) and encryption algorithm.

This Proposed work having lot of enhanced techniques to preserve the privacy in data publish. Thus the all techniques will preserve the membership disclosure and provide more utility than the related system. The diversity checks in the Mondrian and suppression slicing will ensure that these techniques will satisfy privacy requirement of l-diversity. The completion all the related system we got the actual idea of secrecy view in distributed database. Basically slicing is the important algorithm with all available methodologies like data publication, bucketization and generalization in the proposed database.

In collaborative data publishing with partitioned data across multiple data providers, each contributing a subset of records di. A data provider could be the data owner itself who is contributing its own records. Requirement is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties, considering different types of malicious users and information they can use in attacks. In Privacy for collaborative data publishing, focus is on insider attack by colluding data provider who can use their own data record to understand the data records shared by other data record providers. This problem can be resolved by using different approaches as m-privacy, Heuristic algorithms, Data provider aware anonymization Algorithm and SMC/TTP protocols.

M-Privacy [17] protects anonymized data against m-adversary (is a situation where data providers are using combination of data for breaching the anonymized records) with respect to given privacy constraint. M-Privacy can also be guaranteed when there are duplicate records; it also includes syntactic privacy constraint, differential privacy constraint and monotonicity of privacy constraints. M-privacy verification: Binary m-Privacy verification algorithm, Top-Down and Bottom-Up algorithms are used for this. This verification process first analyze the problem by modeling adversary space and using heuristic algorithms with effective pruning strategy and adaptive ordering techniques for effectively checking m-privacy with respect to equivalence group monotonicity constraints.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

b. *Architecture of Proposed System:*

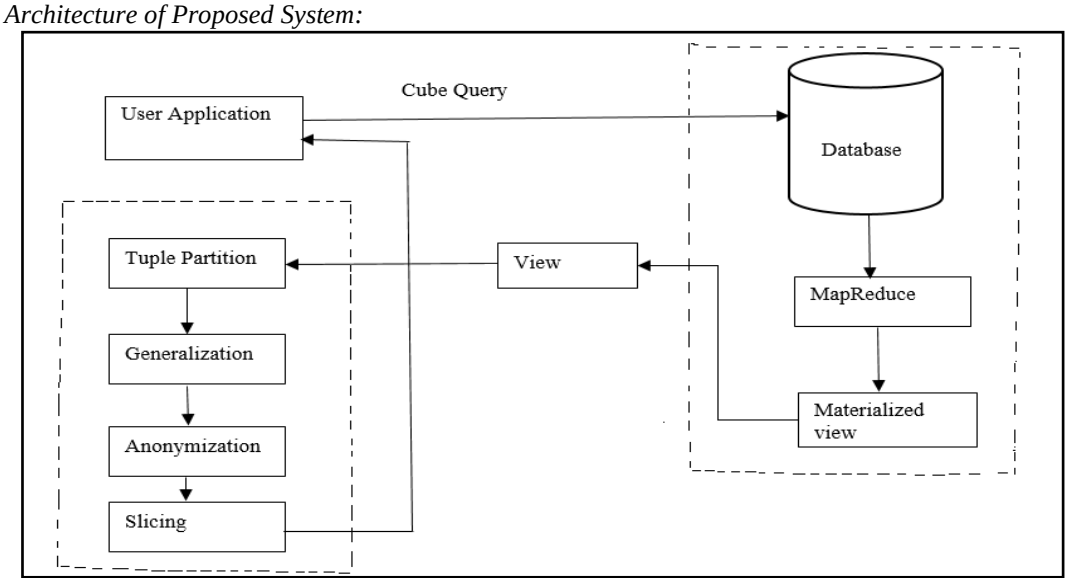

Fig 1: Proposed System Architecture

In Fig.1 shows the architecture of proposed system which contains user module, cube query, MapReduce, Anonymization modules. In first module the user can be an administrator, guest user or authorized one which already have an account. Administrator has all writes to accept request for create account, delete account and give rights. User can access the data on role based. If the user is guest user he/she can access the anonymized data for privacy preservation. Here we are going to use Hadoop framework for managing large scale distributed data and processing it. User passes a query as input in Hadoop framework where data node is master node which receives this query. Cube query is generated annotated lattice which is further given to process main MR-Cube MapReduce Process. The output gets in materialized view after performing MapReduce job on which the anonymization technique is used to generate anonymous data for user.

c. *Description of the Terms And Proposed Algorithm:*

1. Anonymization

Slicing is basically depends on Attribute and tuple partitioning. In Attribute partitioning (vertical partition) we partitioned data as name, age-zip and Disease and tuple partitioning (horizontal partition) as t1, t2, t3, t4, t5, t6. In attribute partitioning age and zip are partitioned together because they both are highly correlated because they are quasi identifiers (QI). These QI can be known to attacker. While tuple partitioning system should check L diversity for the sensitive attribute (SA) column. Algorithm runs are as follows.

Step1. Initialize bucket k=n, int i= rowcount, column count=C, Q=D, // D= data into database, Arraylist= a[i];
Step2. While Q is not empty If i≤n Check L diversity; Else i++; Return D*;
Step3. Q = Q -(D* + a[i]);
Step4. Repeat step 2 and 3 with next tuple in Q
Step5. D* =D * U A[D] // next anonymized view of data D

First initialize k = limit of data anonymization bucket size, number of rows, number of columns, array list and database in the queue(step 1). Further process will done if and only if queue is not empty i.e there should be data in database. Check data for L diversity if rowcount = k = m (step 2). Initially Q= Queue of data. If our bucket data fulfill k anonymity and L diversity, it return D* i.e anonymized view of data. The data from the database which cannot fulfill requirement of privacy will stored in arraylist a[i]. Now data remains in database i.e in Q=Q-D*+a[i] (step3). Repeat step 2 and step 3. A[D] is anonymization of data in database. Apply above steps for remaining data and create new anonymization view which is the union of original view and new one i.e D*=D*UA[D].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

2. *Ldiversity*

Ldiversity is the concept of maintaining uniqueness within data. In this system we used this concept on SA (Sensitive Attribute) i.e on disease. Our anonymized bucket size is 6 and I maintain L=4 i.e from 6 disease record 4 must be unique.

Step1. Initialize L=m, int i;
Step2. If i= n-m+1; Then a[0]..a[1], insert these values as they are in Q; i++;
Step3. Else Check privacy constraint for every incremented value in Q If L=n then Fscore=1 Insert value in the row i++; else Add element to arraylist a[i];
Step4. Exit

First initialize L=m and rowcount i. If i=n-m+1 i.e if k=n=6 and L=m=4 then i=3, upto third row data doesn't need to check for Fscore. Add this data as they are coming from Q(step 1 and 2). For further data from Q check data for privacy constraint. If data fulfills L , then Fscore=1. If data doesnt fulfill Fscore=1, then add element in array list a[i] (step 3).

3. *Permutation*

Permutation means rearrangement of records of data. Permutation process is used for re-arrangement of quasi identifier i.e Zip-Age

4. *Fscore*

Fscore is privacy fitness score i.e the level of fulfillment of privacy constraint C. If fscore=1 then C(D*)= true.

5. *ConstraintC*

C is a privacy constraint in which D* should fulfill slicing condition with L diversity as explain above. Consider value of L diversity is 4. Fscore should be 1 when system fulfills L diversity condition.

6. *Some verification processes are carried out are:*

A. *Verification for L diversity:*

For verification of L diversity I used Fitness score function. For checking L diversity generate continuous similar values of SA i.e insert similar disease. Check for Fscore=1. If L=m, return Fscore. If privacy breach i.e if anonymized view take data as insertad then it breached privacy. D* shoud take data which fulfill L= m.

1. Generate continuous similar values of SA
2. Check for privacy constaint and fscore=1;
3. If Privacy breach; Then early stop; Else Return (Fscore);
4. Exit

B. Verification for strength of system against number of provider:

For verification against number of provider ,add one more attribute in anonymized data as a provider to output. This verification will prove that our technique of anonymization doesn't depend on number of provider. Existing system i.e provider aware anonymization algorithm depends on database as well as provider.

1. Generate values of SA by providers P= 1..n
2. Check for privacy constraint and Fscore=1 with respect to number of provider
3. If Privacy breach; Then stop; Else Return(Fscore);
4. Exit

IV. PSEUDO CODE

PROVIDER AWARE ALGORITHM FOR REDUCE THE TIME COMPLEXITY:

Input: Data set with D, providers n, with C
Output: : Slice view (T∗) with provider
1. read data from (D up to null)
2. for each (attributes in table) for each (tupels in tables)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

3. Set quasi identifier (QIfr) and sensitive attributes (SA)
4. Apply generalization technique it will classify the tuples in QIfr groups
5. Apply anonymization on relative information attributes
6. While(verify data-privacy(D, n, C) = 0) do if (Di→D) verified with QIfr then add Di up to when K-anonimty else ealy stop Bucket(i1)→D; 7. permute the data with (I=(I(null-1)))
8. Apply Pruning on (D)
9. Apply step 1, 2, 3 on Bucket(i1)
10. if (C fails with (D)&&(p#1)) Bucket(i2)→Bucket(i1(j))
11. Display all (Bucket (i2)6=null) 12. end while
13. end for

V. MATHEMATICAL MODEL FOR PROPOSED SYSTEM

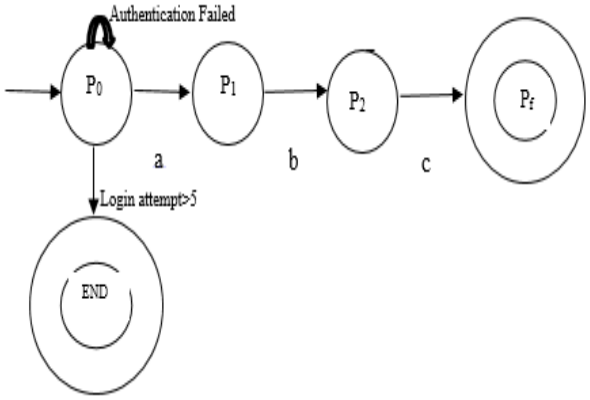

Fig 2: DFA of Proposed System

DFA={Q, ∑, δ, q0, F}
Where
Q=Finite Set of States
∑=Input Alphabet
δ=Transition between states
P0=Initial State
F=Final State

Q={P0, P1, P2, Pf}
P0=Initial State
P1= Create Cube Query
P2=MapReduce
Pf=Anonymization

∑={a, b, c} Where
a=Query with parameter
b=annotated lattice
c=Materialized data

	a	b	c
P_0	P_1	Φ	Φ
P_1	Φ	P_2	Φ
P_2	Φ	Φ	P_f
P_f	Φ	Φ	Φ

Table 1: State Transition table

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

VI. EXPERIMENTAL RESULTS

We present here sets of experimental results to 1) compare and evaluate query processing time in Hive and using cube 2) evaluate and compare proposed updated provider aware algorithm for given dataset to get more data utility with securing data.

a. Experiment Setup:

We used healthcare dataset which contains different attribute like name, age, zip, address, disease etc. 1 Lacks of records have been used in all experiment. The Disease have been used as a sensitive attribute (SA). This attribute has 10 distinct values. Data are distributed among 4 providers p1, p2, p3, p4. The privacy constraint C is defined by k-anonymity and l-diversity. C is conjunction of both k-anonymity and l-diversity. Anonymization use Fscore i.e. privacy fitness score, if the diversity is 3 the fitness score is 1, for diversity 5 the Fscore will 2.

b. Query Processing with cube:

We used Hadoop Apache open source framework for storing and processing data. Hive is used as a sql in Hadoop. We generate a patient cube with four dimensions (name, disease, age ,doctor) and two measures (provider and zip) We compare time required to processed query in Hive and using MR-Cube. In fig.1 shows the comparison which shows building a cube lattice over the dataset we can retrieve data in less time with more accuracy than time required to data retrieving through Hive. Fig. 2 shows the performance of data insertion in proposed system. In existing encryption based system required large amount of time to insert data as it used very lengthy process.

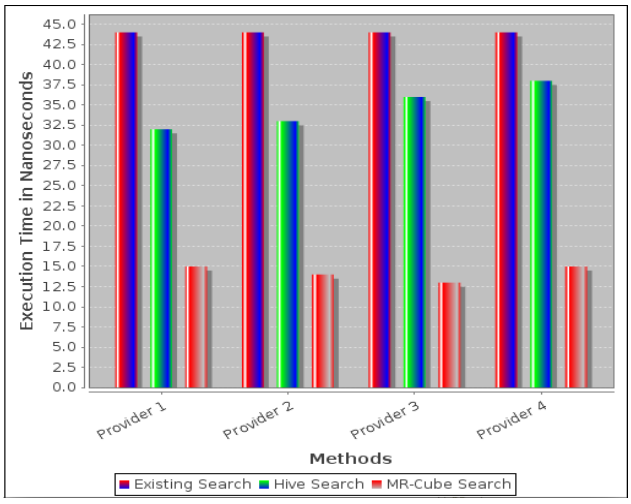

Fig.1 Data Extraction Using Different Methods

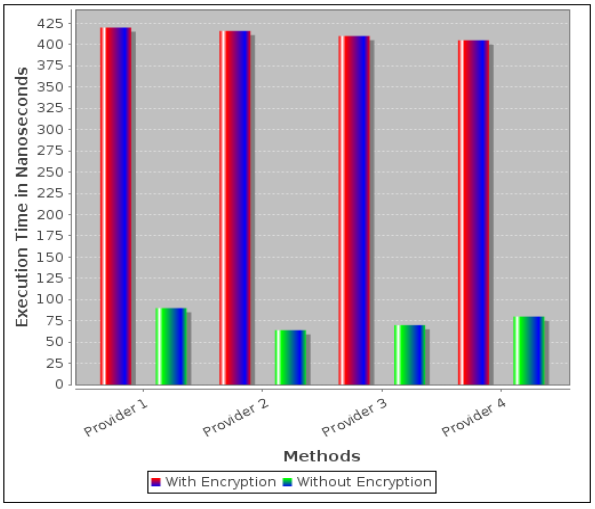

Fig. 2 Data Insertion performance

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

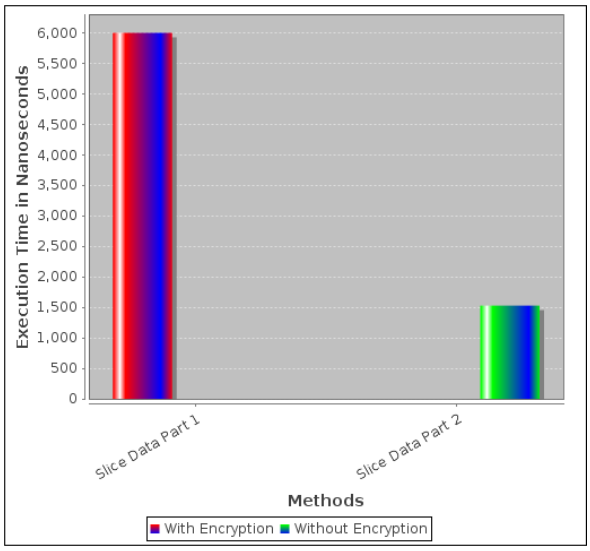

Fig. 3 Data Slicing Performance

VII. CONCLUSION AND FUTURE WORK

The results showed that the proposed algorithm performs better than existing one. The more data get utilized for analysis without breaching privacy of individual. Existing system uses encryption before inserting/ storing data and decryption using key to retrieve it. It required large amount of time and it was not worked efficiently for big data. Using MR-Cube approach data cube can compute efficiently. The cube has been generated for given dataset with dimensions and measures. MR-Cube compute cube with holistic measures like Top-k query so get accuracy. The proposed algorithm uses to reduce time complexity as existing system uses multiple checks for privacy constraint.

REFERENCES

1. Shital Suryawanshi, Prof. V.S.Wadne,' Big Data Mining using Map Reduce: A Survey Paper', IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. VII (Nov – Dec. 2014), PP 37-40, 2014.
2. Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE "Data Mining with Big Data, "in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.
3. Tiancheng Li, Ninghui Li, Jian Zhang, Ian molloy "Slicing: A New Approach for Privacy Preserving Data Publishing "in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012.
4. Madhuri Patil, Sandip Ingale "Privacy Control Methods for Anonymous and Confidential Database Using Advance Encryption Standard "in International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 8, August 2013.
5. Senthil Raja M and Vidya Bharathi D "Enhancement of Privacy Preservation in Slicing Approach Using Identity Disclosure Protection "in ITSI Transactions on Electrical and Electronics Engineering (ITSITEEE) Volume -1, Issue -2, 2013.
6. HArnab Nandi, Cong Yu, Philip Bohannon, and Raghu Ramakrishnan, Fellow, IEEE, "Data Cube Materialization and Mining over MapReduce "TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 6, NO. 1, JANUARY 2012..
7. Zhengkui Wang, Yan Chu, Kian-Lee Tan, Divyakant Agrawal, Amr EI Abbadi, Xiaolong Xu, "Scalable Data Cube Analysis over Big Data "appliarXiv:1311.5663v1 [cs.DB] 22 Nov 2013.
8. ArnabNandi, CongYu, PhilBohannon,RaghuRamakrishnan"DistributedCubeMaterialization on Holistic Measures "
9. Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, Member, IEEE "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud "in IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 2, FEBRUARY 2014.
10. SlawomirGoryczkLiXiongEmory,BenjaminC.M.Fung"m-PrivacyforCollaborativeData Publishing ".
11. Ashwin Machanavajjhala, Johannes Gehrke, Danial Kifer "-Diversity: Privacy Beyond kAnonymity ".
12. Dhanshri S. Lad , Rasika P. Saste, "Different Cube Computation Approaches: Survey Paper"(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4057-4061
13. K. V. Shvachko and A.C. Murthy, "Scaling Hadoop to 4000 Nodes at Yahoo"Yahoo! Developer Network Blog, 2008.
14. NOMAN MOHAMMED and BENJAMIN C. M. FUNG, PATRICK C. K. HUNG, CHEUKKWONG LEE, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data "in ACM Transactions on Knowledge Discovery from Data, Vol. 4, No. 4, Article 18, Pub.date: October 2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

15. C.Aggarwal, "On K-Anonymity and the Cure of Dimensionality "Proc. Intl Conf. Very Large data Bases (VLDB), PP, 901-909, 2005.
16. D.Mohanapriya,Dr.T.Meyyappan,"HighDimensionalDataHandlingTechniqueUsingOverlapping Slicing Method for Privacy Preservation "in International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 6, June 2013 .
17. BenjaminC.M.Fung,KeWang,andPhilipS.Yu,Fellow,IEEE,"AnonymizingClassification Data for Privacy Preservation "in EEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 5, MAY 2007.
18. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data, "ACM Crossroads, vol. 19, no. 1, pp. 2023, 2012.
19. K. V. Shvachko and A.C. Murthy, "Scaling Hadoop to 4000 Nodes at Yahoo "Yahoo! Developer Network Blog, 2008.
20. Hadoop. "http://hadoop.apache.org/ ".
21. TheApacheSoftwareFoundationhttp://hadoop.apache.org/docs/current/hadoopyarn/hadoopyarn-site/YARN.html