Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

A Secure Big D ata Framework Based on Access Restriction and Preserved Level of Privacy

Folorunso O.

Received: 9 December 2019 Accepted: 5 January 2020 Published: 15 January 2020

Abstract

Big data frequently contains huge amounts of personal identifiable information and therefore the protection of user?s privacy becomes a challenge. Lots of researches had been administered on securing big data, but still limited in efficient privacy management and data sensitivity. This study designed a big data framework named Big Data-ARpM that is secured and enforces privacy and access restriction level. The internal components of Big Data-ARpM consists of six modules. Data Pre-processor which contains a data cleaning component that checks each entity of the data for conformity.

Index terms— differential privacy, big data, access restriction, data privacy

1 Introduction

he developing wonder called big data is compelling various changes in organizations and different associations. Many battle just to deal with the gigantic informational collections and nonconventional information structures that are commonplace of big data.

Large information management is about two concepts: big data and data management, plus how the two work together to accomplish business and innovation objectives.

According to Ray (2018) Big Data refers to a large volume of diverse, complex and fast-changing data, derived from new data sources. The data sets are so large that is very difficult to manage by the traditional data processing software or the traditional software management (Manyika et al., 2011; **??**ürsakal, 2014).

Big data is first about data volume, namely large datasets measured in tens of terabytes, or sometimes in hundreds of terabytes or pet abytes. Also, big Data is so huge and complex that it is impossible for traditional systems and traditional data warehousing tools to process and work on them. Before the term enormous information became regular speech, we discussed Very Large Databases (VLDBs). VLDBs usually contain exclusively structured data, managed in a database management system (DBMS).

Notwithstanding exceptionally huge datasets, large information can likewise be a mixed blend of organized information (social information), unstructured information (human language content), semi-organized information (RFID, XML), and spilling information (from machines, sensors, Web applications, and web-based social networking). The term multi-organized information alludes to informational collections or information conditions that incorporate a blend of these information types and structures. (Gantz and Reinsel, 2011).

2 T

With the expansion in the utilization of big data in business, numerous organizations are grappling with privacy issues. Information protection is a risk, consequently organizations must be on security cautious. Security is the case of people, gatherings, or organizations to decide for themselves when, how, and to what degree data about them is imparted to other people. In contrast to security, privacy ought to be considered as a benefit; in this manner it turns into a selling point for the two clients and different partners. There ought to be a harmony between data privacy and national security.

3 II.

4 Related Work

Lu et al., (2014) made a methodology towards the proficient and protection saving processing in the big data period, and it misuses the new difficulties of big data in security safeguarding. At first, it characterizes the general engineering of big data examination and finds the security necessities in big data. At that point, it discovers a proficient and privacy preserving cosine closeness figuring convention. The limitation of the work is that it needs significant research efforts for addressing unique privacy issues in some specific big data analytics.

Xu et al., (2016) structured a system named "Rampart framework" for privacy safeguarding. It comprises of techniques in particular anonymization, recreation, change, provenance, understanding, exchange and limitation to forestall outside interruption. The system endeavored to give high need to keep up the harmony between information utility and privacy however recommended that more ways are to be investigated to ensure protection against different dangers.

Shrivastva et al., (2014) broke down how much the differential privacy approach is appropriate for big data security conservation and introduced different elements that assume key job in big data security safeguarding. Among the different methodologies, differential privacy is the best appropriate for big data as it is liberated from the imperfections of different methodologies. Plus, differential privacy looks for balance among utility and security. A framework of perturbation is introduced to accomplish the differential privacy.

Al-Aqeeli and Alinfie (2015) researched some security saving issues of big data with regards to half breed distributed computing and assessed a few systems, for example, Airavat, Sedic, Sac FRAPP and Hyper-1 dependent on Map Reduce from the point of view of versatility, cost and similarity. It was recorded that anonymization, encryption, differential privacy are the productive strategies for ensuring protection of information. The last investigation shows that the featured structures experiences constraints, for example, information contortion and none of them is completely fit for privacy preservation. Mehmood et al., (2016) introduced existing protection safeguarding instruments in the different life patterns of big data, for example, data generation (encryption and access limitations), information stockpiling (hybrid and private mists) and information handling (generalization, suppression, anatomization, permutation and perturbation) and different difficulties of saving security in large information. These techniques were portrayed as for the variables of versatility, security, time, proficiency and utility. Different dangers engaged with the encryption, anonymization and capacity of information in the cloud were likewise researched. At the point when these strategies are applied, security is ensured however the information may lose the importance in reality and thus the utility and criticalness. For information distributing, a calculation must consider legitimate exchange off among utility and security as the information is inclined to any assaults. Along these lines the strategies/methods must be adjusted or stretched out to deal with the large information in a proficient way. Yan et al., (2016) proposed a pragmatic plan to deal with the encoded big data in cloud with de duplication dependent on possession challenge and Proxy Re-Encryption (PRE). As recognized by Jian et al., (2016), the constraint of their work is that Convergent Encryption (CE) is dependent upon an innate security restriction for example powerlessness to disconnected animal power word reference assault.

Sedayao et al., (2014) introduced a contextual investigation of anonymization in an endeavor identifying the necessities and execution detail for saving security of enormous information. Anonymized informational collections must be painstakingly broke down, estimated and tried whether they are inclined to any assaults since it is more than covering or generalization. The creators recommended the utilization of Hadoop to break down and get helpful outcomes from the big data. The analyses are led with static informational index, yet it ought to be stretched out for continuous informational indexes. The work couldn't reason that the anonymized information is completely liberated from any sort of assaults.

Zakerdah and Aggarwal (2015) proposed a methodology towards protection safeguarding information mining of exceptionally monstrous informational indexes utilizing map reduce. They study two most broadly utilized security models k-anonymity and l-diversity variety for anonymization, and present test results outlining the effectiveness of the methodology. The constraint of their work is that generalization cannot deal with high dimensional information, it decreases information utility. Perturbation decreases utility of information. Zhang et al., (2013) proposed Cloud Safe to redesign availability and mystery of the set aside information in the cloud through scrambling and encoding data into a couple of disseminated stockpiling providers. Cloud Safe offers a cloud-based individual electronic asset safe help which passes on the critical assets between a couple of cloud providers by using destruction coding and cryptography. As per Zhang et al., (2013), the accessibility improves because of utilizing eradication coding to disperse the information on a few cloud suppliers, so as to recoup information get to when a supplier falls flat.AES was utilized for scrambling and unscrambling information to keep information secrecy.

Zhang et al., (2014) researched the versatility issue of multidimensional anonymization over big data on cloud, and proposed an adaptable Map Reduce based methodology. The flexibility issues of finding the center on account of its inside activity in multidimensional allotting was investigated and significantly versatile Map Reduce based computation was proposed for finding the center and histogram strategy. Logically number of investigations on datasets were coordinated which was removed from real datasets, and the exploratory results show that the flexibility and cost sufficiency of multidimensional anonymization plan can be improved basically over existing

techniques, anyway ensuring insurance protecting of immense extension educational assortments regardless of everything needs wide assessment.

Pramanik et al., (2016) presented a conceptual framework that integrates and improves technologies for preserving big data privacy. The proposed model empowers the structure of a dependable protection framework for a given e-government procedure and comprises of three significant modules: a) Big information assortment, b) Information extraction, and c) Anonymization module. In this work, a Conditional Random Field (CRF) classifier was conveyed for extricating distinguishing characteristics, and kanonymization strategy for de-recognizing the separated information through insignificant speculation and concealment. The creators likewise introduced a lot of primer trial results indicating the viability of the proposed structure dependent on some security assessment measurements.

5 III.

6 Design Methodology

The architecture named Big Data-ARpM (Big Data Access Restriction and Privacy Mechanism) is defined by the collection or gathering of data with high velocity, volume and different varieties, classification of the gathered data, storing the data securely and restricting the access to data from within and out of the systems. Figure 1a is a physical architecture that gives an insight into the operational structure of Big Data-ARpM, Figure 1b shows the internal structures of the Access Restriction and the Key Management Module, while Figure 1c shows the internal structure of the request management module. The architecture is designed to run on a distributed server environment and to store and retrieve data from a parallel database system; this is because of the high velocity, volume and different varieties of data. Big Data-Arp Mretrieves input from synchronous multiple data sources, these input are raw and however need to be pre-processed and further classified before its then stored securely and await a request for delivery, because the data may contain sensitive information of so many entities, people and organization, releasing the data without anonymizing them may be a very great disaster Big , Data-Arp Mhas a well-structured internal component that facilitates all the processes, the structures and their respective functions are;

7 b) Data Pre-processor Module

Data pre-processing is an important step in data gathering, data gathering are mostly loosely controlled, resulting in out of range value (Age: -100) and impossible data combinations e.g. (Sex: Male, Pregnant: Yes). Data that are being gathered and input from the source (WebCrawler) are considered to be noisy-data, however the Data Pre-Processor Module contain a data cleaning component that check each entity of the data for conformity, the output from the data pre-process is a processed and filtered data.

8 Data Classification Module

This module deals with the classification of data due to the sensitivity of such data. The role assigned to the user will determine what class of data such data users can access. There are three basic levels of classifications in this module, which are:

? Normal Level: Users assigned to this level can only view attributes such as the Quasi-identifier (QID).

(QID) is a set of attributes such as zip code, gender, a birth date in which the combination of this attributes could potentially distinguish individuals. This level is the least sensitive of all the three levels.

9 Data Preservation Module

The Module consist of two sub modules with the goal of preserving data before release to any user or any third party application to prevent the privacy violation of the data owner. The data passes through the first module that build an aggregated tree of a single sink data from various data coming from various sources of data entries; this reduces the chances of tracing back the data back to the original owner , prim's algorithm was employed to build the tree. The aggregated data is then passed on to the differential privacy module, which introduce a minimum distortion in the information provided by the database system.

10 ALGORITHM 3: Differential Privacy Algorithm Input: Level, dp Request Output: DP_response Begin

Step 1: The analyst can make query to the database through this intermediary privacy guard.

Step 2: The privacy guard takes the query from the analyst and evaluate this query and other earlier queries for the privacy risk. After evaluation of the privacy risk.

Step 3: The privacy guard then gets the answer from the database Step 4: Add some distortion to it according to the evaluated privacy risk and finally provide it to the analyst.

11 End a) Access Restriction and Key Management module

This modules consists of different sub modules, that coordinates the user or third party application registration, access to data and information in the entire systems, the modules are Due to high velocity and large volume of data that will be passing through the system, this module is designed to handle all split processes in parallel across a cluster of servers and also store and retrieve data across a distributed storage devices, the modules uses Map Reduce, which is programming model for processing large set of data with a parallel and distributed algorithm across a cluster of server.

12 c) Request Management Module

The module handles all incoming request from either application users and or third party applications with the aid of the access restriction module which verify the membership of the users, and also analyse the request to know the level of information been requested, check if the level of the user can access the level of information requested. After the user successful verification, the users query/ request passes through differential privacy technique which deny the users direct access to the database.

13 Data Set

A medical dataset was used in the implementation of Big Data-ARpM, the dataset, named Health Care Provider Credential Data was downloaded from an open source called "data.wa.gov". The dataset contains more than a million instances (records) and 12 attributes (Columns). The computation time is measured in milliseconds and on Big Data platform, the comparisons depict that DP takes lesser computational time in protecting data privacy against k-Anonymity to complete protecting data privacy with 0.4 and 0.45 milliseconds values respectively. This shows that DP is quiet better when privacy protection of data is needed and processing time is to be considered in Big Data analytics. This illustrate that DP produced more records that is useful for analyst with an increase in privacy level while as the privacy level (k) increased, k-Anonymity produced few records with the utility lesser than DP. For instance, when the privacy level applied is 20 and 60, DP present a total of 100 and 300 records against 40 and 190 records produced by k-Anonymity which shows that as the level of privacy level in DP generate more useful records that can be used for analysis while the confidentiality of data are hiding. Though, DP and k-Anonymity have the same privacy level (k), the utility of records generated from each algorithm differs and depict that DP produced more useful data than k-Anonymity.

14 VII.

15 Results and Discussions

16 Figure 7: Comparison between the noise, privacy and utility levels

In this paper, the approach DP used applied noise variant in achieving its purpose as depicted in Figure **??** showing the comparison between the noise, privacy and utility levels. The privacy level shows the protection of data from being identified, the utility level shows the usefulness of data after nose has been added to user's queries and noise level is privacy balanced added to individual record based on the attribute of each data and the level of each user requesting for data. For example, when the noise level is 10, privacy and utility level are 20 and 95 values respectively revealing that the more noise added, there is increase in the privacy and decrease in the utility of data information presented to the users. This also shows that, there is every possibility that we have the same level of privacy and utility of data as shown where we have noise level to be 35 added, privacy and utility levels having 60 respectively and this means that DP +Noise give rise to privacy preserving of data with a reasonable amount of utility than k-Anonymity algorithm.

17 VIII.

18 Conclusion

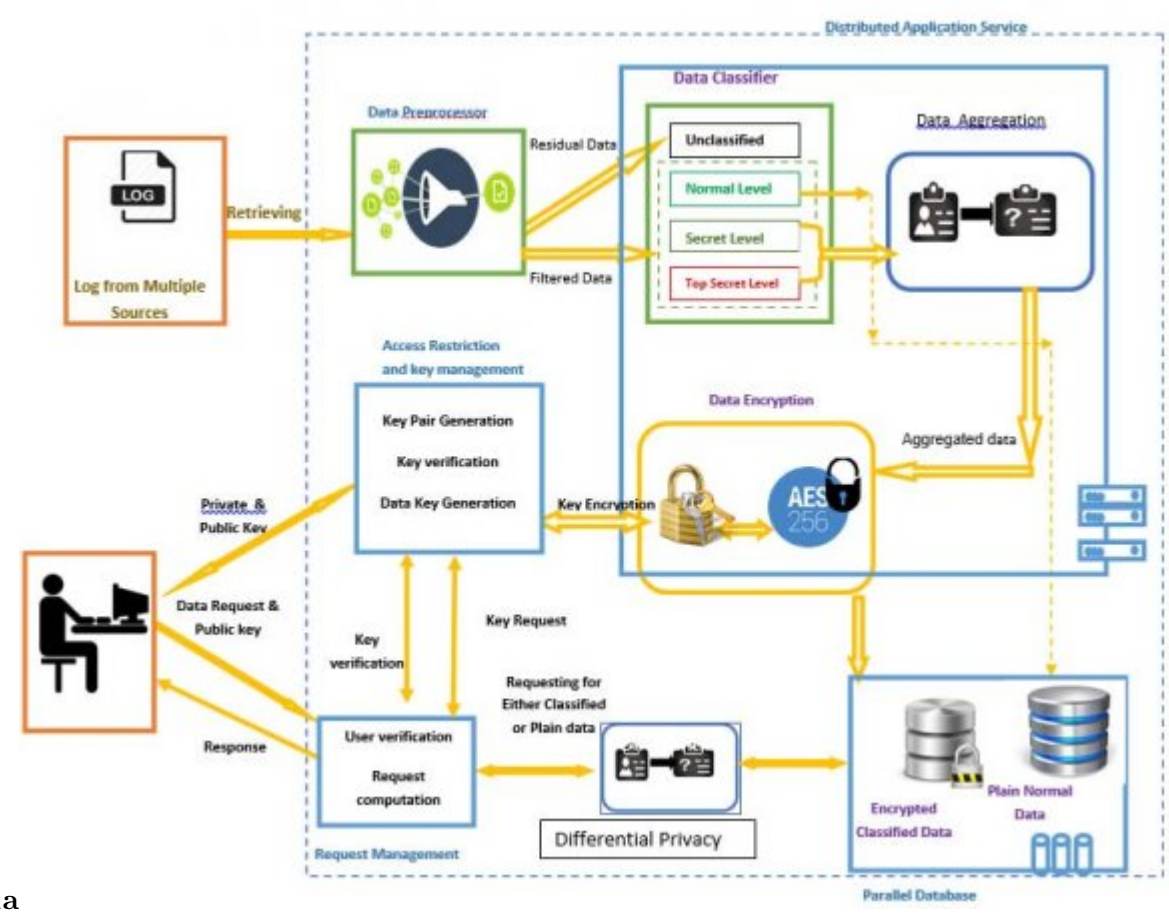
In this study, a conceptual privacy and access restricted framework for securing big data was conceived by designing a data classification scheme according to degree of confidentiality and also designing a privacy preservation technique that enforces data privacy based on data aggregation and differential privacy.

Conclusively, Big Data-ARpM was evaluated based on its utility, scalability, accuracy, sensitivity, specificity and processing time. The results shows that Big Data-ARpM has a very good utility, highly scalable, accuracy of 95.80%, sensitivity of 93.60%, specificity of 98.00% and an execution time of 0.4 milliseconds, as compared with other privacy preservation techniques such as K-anonymity. Hence, the usage of differential privacy technique in Big Data ARpM show that the framework is far better than other frameworks that makes use of other technique.

19 IX.

20 Recommendation

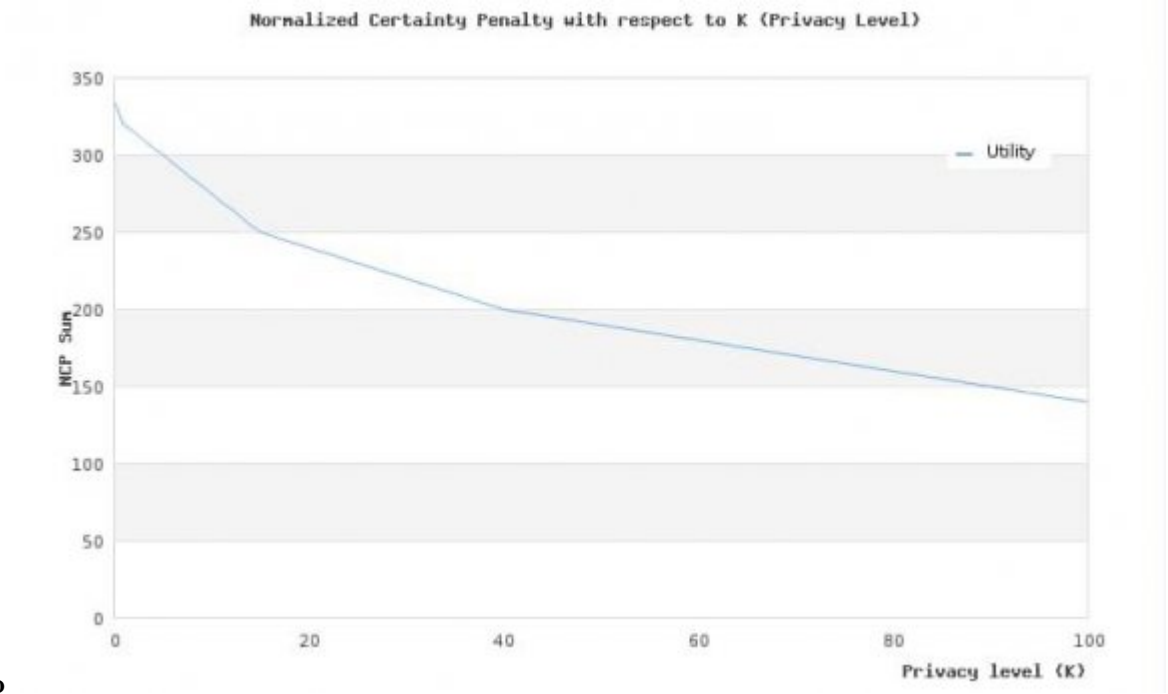
With the efficient techniques presented in this research work, it is believed that the study can be easily extended to focus more on other type of data such as the semi-structured data and unstructured data. Finally, the presented framework can be built upon to accept larger files of different formats. [1]



1a

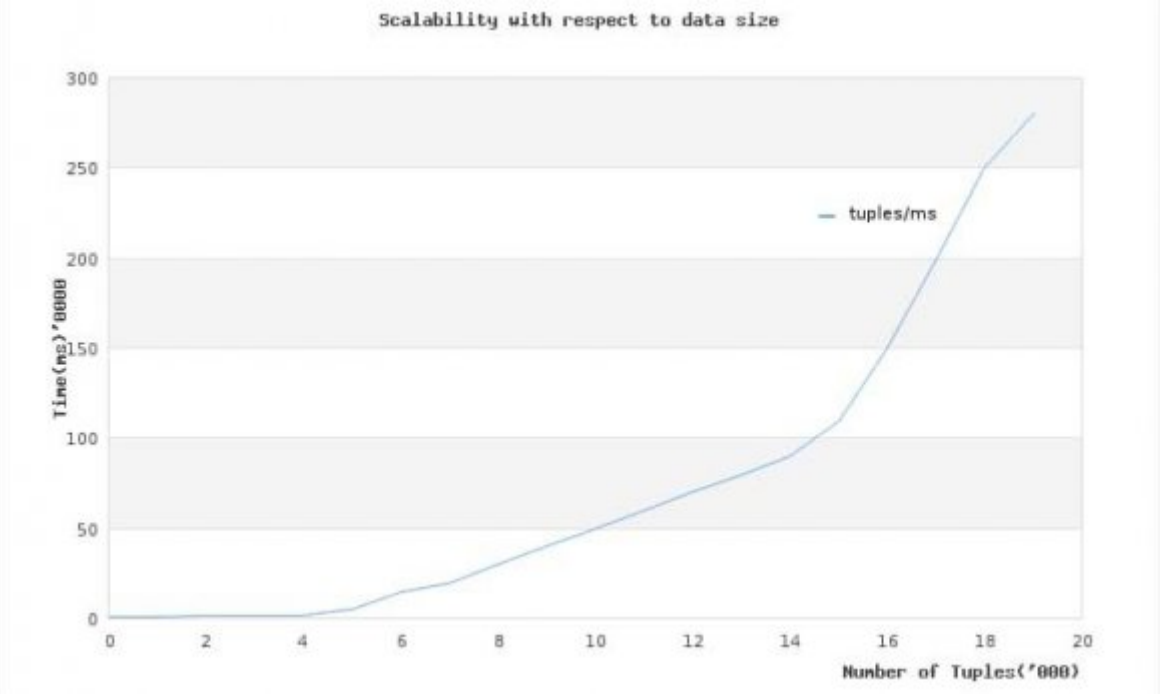
Figure 1: Figure 1a :

[1] © 2020 Global Journals



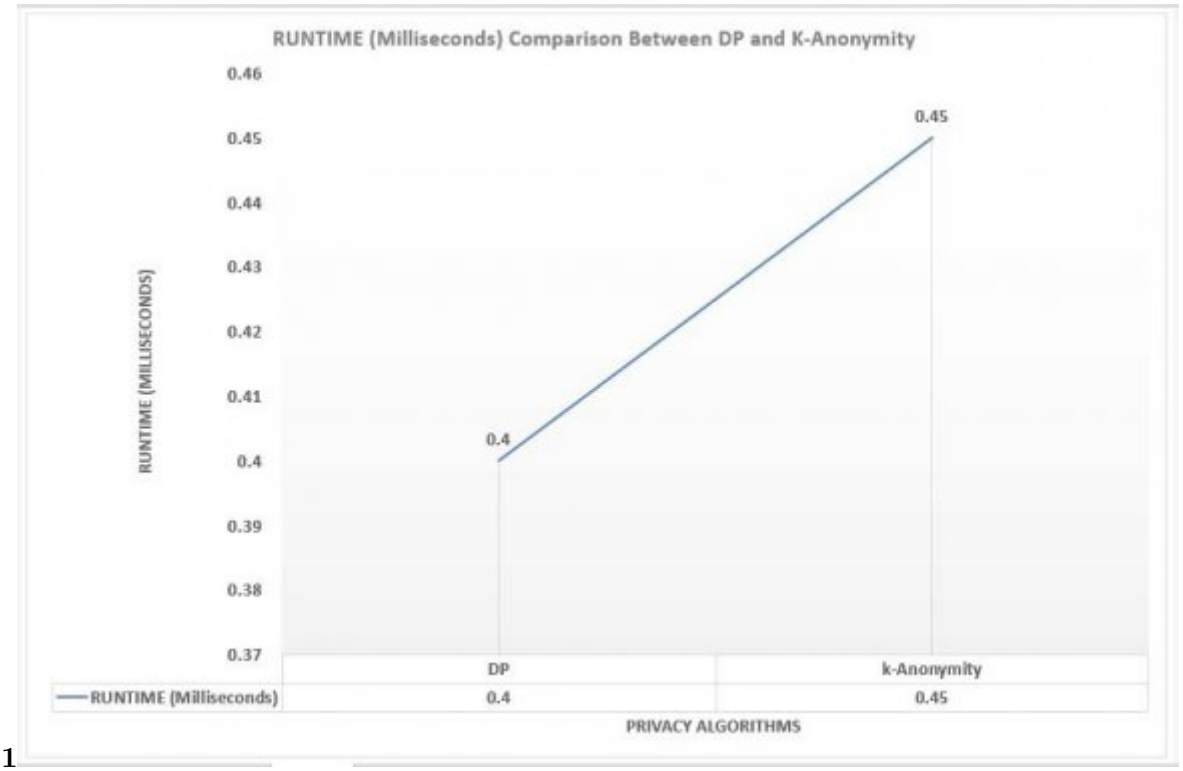
1b

Figure 2: Figure 1b :



1c

Figure 3: Figure 1c :



1

Figure 4: ALGORITHM 1 :

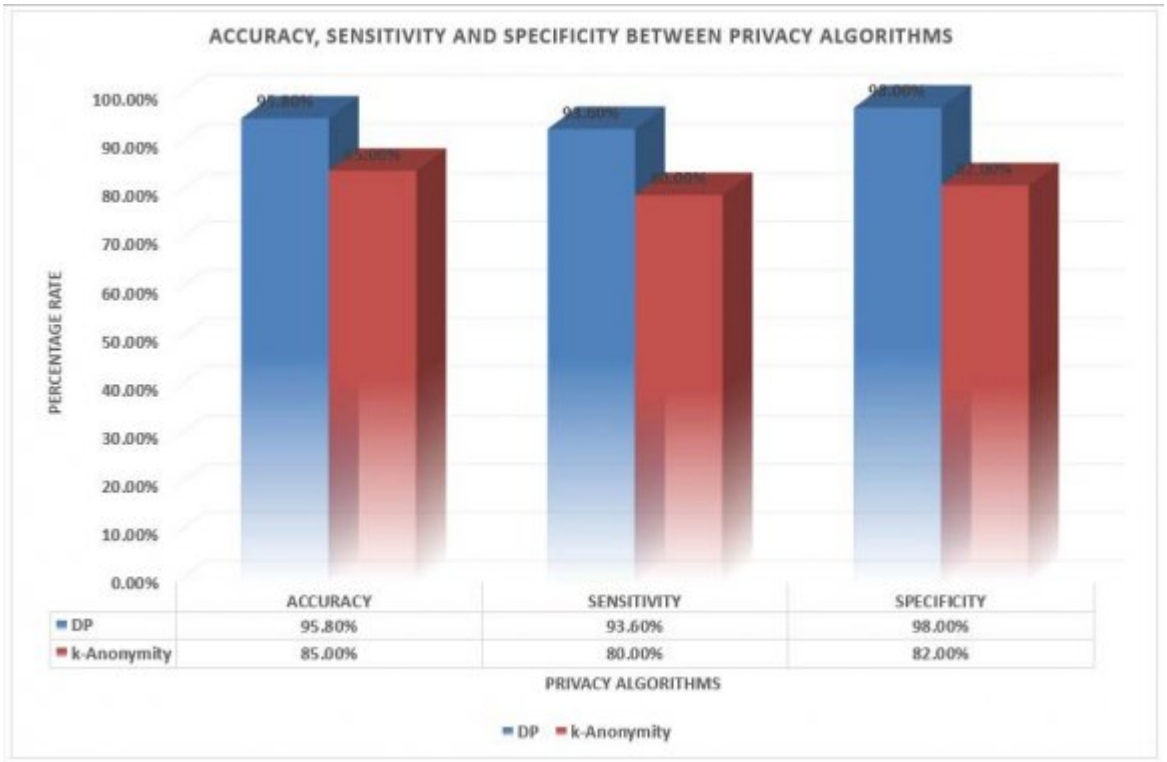


Figure 5: ?E?

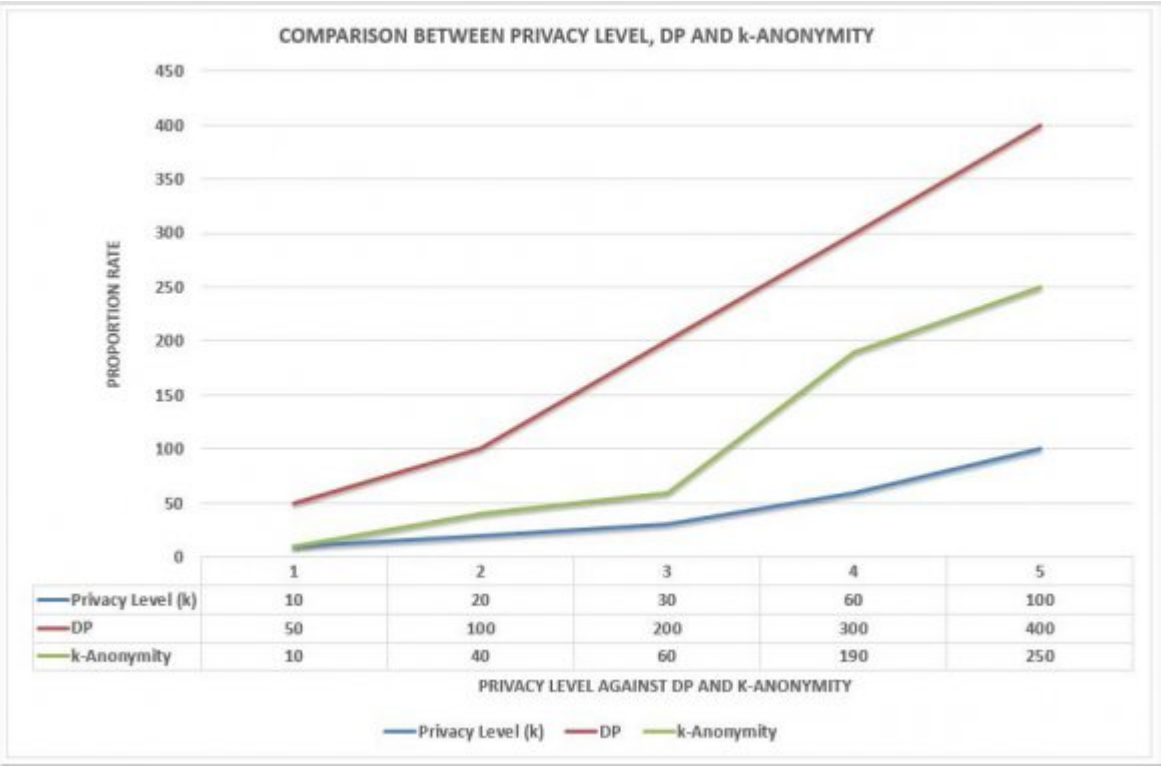
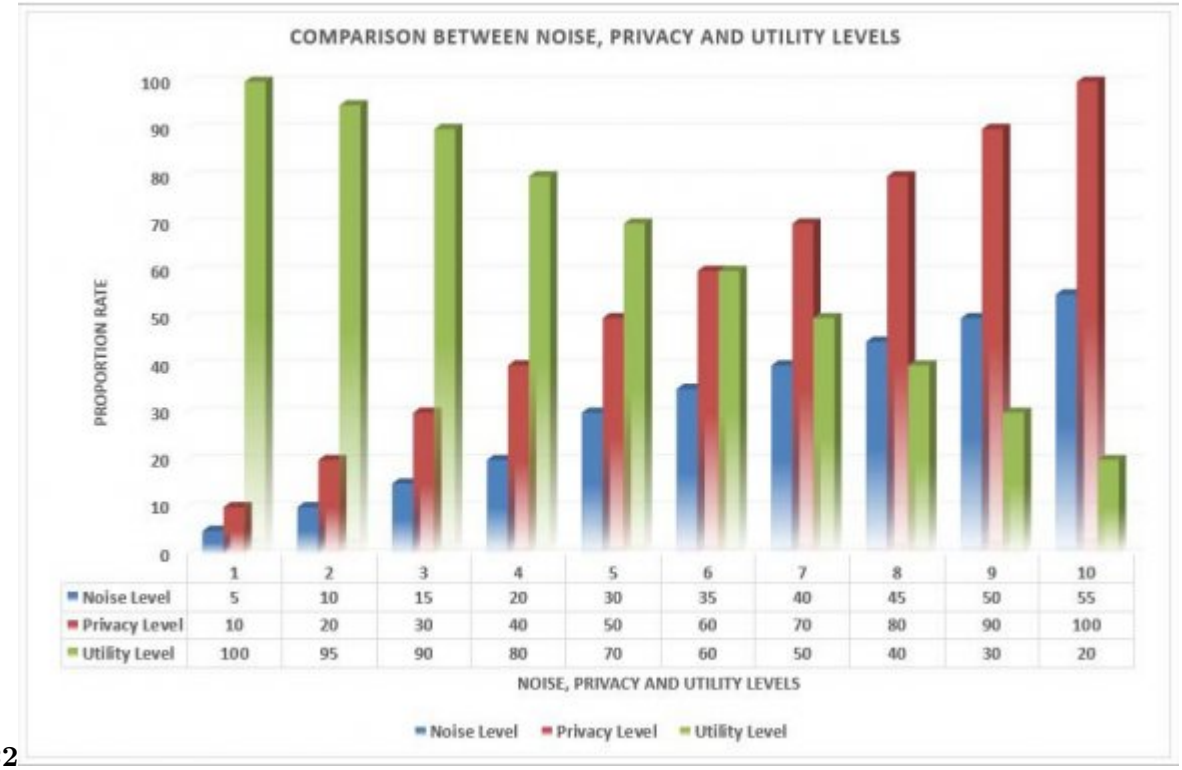


Figure 6: ALGORITHM



22

Figure 7: Figure 2 :Figure 2

Input: Incoming request.
Output: Preserved outgoing data
Begin
Step 1: Login credentials validated by access restriction module ?True/False
Step 2: If "True", request interface is displayed. Access Granted.
Otherwise, the user is an unauthorized user. Access Denied.
Step 3: If "Access Granted" Then
Level ? call Request User Level();
Send Request (req, level, res);
dp? DP(res,level);
Step 4: Is True (dp): process Result (dp?result):
preserved Data (dp?result);
Step 5: Output Request (dp?result);
Step 6: otherwise, goto step 3.
End
VI.

Figure 8:

1

Year 2020
70
Volume XX Issue III Version
I
( ) E
Global Journal of Computer Science and Technology

Evaluation Metrics	Differential Privacy (DP)	K-Anonymity
Data Utility	High Good	Low Poor
Scalability Accuracy	95.80% 93.	85.00%
Sensitivity Specificity	60% 98.00% 0.40	80.00%
Processing Time	ms	82.00% 0.45 ms

Figure 9: Table 1 :

[Xu et al. ()] 'A Framework for Categorizing and Applying Privacy Preservation Techniques in Big Data Mining'. L Xu , C Jiang , Y Chen , J Wang , Y Ren . *Computer* 2016. 49 (2) p. .

[Pramanik et al. ()] *A Privacy Preserving Framework for Big Data in E-Government*, Mdileas ; Pramanik , Lau , Y K Raymond , Wei T Yue . 2016. 2016.

[Shrivastva et al. ()] *Big Data Privacy Based on Differential Privacya Hope for Big Data*, K M Shrivastva , M Rizvi , S Singh . 2014. p. 167.

[Manyika et al. (2011)] *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, J Manyika , M Chui , B Brown , J Bughin , R Dobbs , C Roxburgh , A H Byers . 2011. May 2011. Seattle: McKinsey Global Institute.

[Zhang et al.] *Cloud Safe: Storing Your Digital Asset in the Cloud-based Safe*, Q Zhang , B Luo , W Shi , A M Almoharib . p. . Wayne State University

[Zhang et al. ()] *CloudSafe: Storing Your Digital Asset in the Cloudbased Safe*, Q Zhang , B Luo , W Shi , A M Almoharib . 2013. p. . Wayne State University

[Yan et al. ()] 'Deduplication on encrypted big data in cloud'. Z Yan , W Ding , Xixun Yu , H Zhu , R H Deng . *IEEE Trans Big Data* 2016. 2 (2) p. .

[Gantz and Reinsel ()] 'Extracting value from chaos'. J Gantz , D Reinsel . *IDC iview* 2011. 1142. p. .

[Gartner 2014. IT Glossary, What is Big Data? URL ()] Tarihi: 20.12. http://www.gartner.com/it-glossary/big-data/ *Gartner 2014. IT Glossary, What is Big Data? URL*, 2014. (Son Eri?im)

[Sedayao et al. ()] 'Making Big Data, Privacy, and Anonymization Work Together in the Enterprise: Experiences and Issues'. J Sedayao , R Bhardwaj , N Gorade . 10.1109/bigdata.congress.2014.92. *IEEE International Congress on Big Data*, 2014.

[Al-Aqeeli ()] 'Preserving Privacy in Map Reduce Based Clouds: Insight into Frameworks and Approaches'. S Al-Aqeeli , Alnifie , G . doi:10.110. *International Conference on Cloud Computing (ICCC)*, 2015.

[Zakerdah ()] *Privacypreserving big data publishing*, H C C Zakerdah . 2015. La Jolla: ACM.

[Mehmood et al. ()] 'Protection of big data privacy'. A Mehmood , I Natgunanathan , Y Xiang , G Hua , S Guo . *IEEE translations and content mining are permitted for academic research*, 2016. p. .

[Hashem et al. ()] 'The rise of "big data'. I A T Hashem , I Yaqoob , N B Anuar , S Mokhtar , A Gani , S U Khan . *Review and open research issues. Information* 2015. 47 p. .

[Ray ()] *The-complete-beginners-guide-to-bigdata-in-2018*, R Ray . https://medium.com/swlh/the-complete-beginners-guide-to-big-data-in-201882ed7a396ba3 2018.

[Lu et al. ()] 'Toward efficient and privacy-preserving computing in big data era'. R Lu , H Zhu , X Liu , J K Liu , J Shao . 10.1109/MNET.2014.6863131. https://doi.org/10.1109/MNET.2014.6863131 *IEEE Network* 2014. 28 (4) p. .