International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

A Novel Approach to Ensure Privacy in Big Data

D. Angelin Benita[1], S.P. Victor[2]
[1]*Research Scholar, Department of Computer Science, St. Xavier's College, Palayamkottai, India*
[2]*Department of Computer Science, St. Xavier's College, Palayamkottai, India*

Abstract: ***The latest buzzword in the information technology sphere is Big Data and the technologies pertaining to it. With the greater access to the Internet through the so called Internet of Things (IoT) content exchange through various media via the Internet has massively increased. All the data obtained has to be stored somewhere and with massive quanta of data stored through clouds it gives an opportunity to mine massive amount of Knowledge. Although this is welcome direction it can also invade into privacy and security of a person. If for example consider a hospital which stores massive amounts of data, it can be useful to mine potential patterns of symptoms among various patients for a particular disease spread across various age groups and hence forth. But the data can also reveal details about a single person. If this info falls into the hands of the wrong person then we are talking about intrusion of privacy. Security and Privacy are two different terms. Security determines who can access the data which can be monitored and privacy becomes an issue when what one does with that data. Lot of security mechanisms are in place nowadays. But privacy is still a wanting topic. Hence in this approach a unique approach is being proposed that will safe guard one private information is the age of Big Data.***

Keywords*: *Big Data, Security, Internet of Things.

I. INTRODUCTION

The massive growth of the internet over the past two decade coupled with latest technologies like smartphones and the Internet of Things has massively increased the quanta of data being transferred throughout the data. These data are impossible to store in traditional and client storage devices, prompting companies to implement new technologies like cloud computing to store massive amounts of data. Cloud computing is a term used to describe and IT model where pools of resources can be configured (such as servers, networks, services etc) can be configured to store massive amounts of data thereby saving a company from cost incurred in saving huge amounts of data. It allows data to be stored and processed via a third party server which is located in a data center. [1]-[3]The Internet of Things is the network of physical devices that provides data for the cloud. unlike the time during the dawn of the internet when data was coming from only desktop computers, nowadays there are a lot of devices like embedded electronic, sensors, actuators, logs etc which load data into the internet. These actually help in the smooth functioning of many an application and hence are so vital. [4]. The relation between the cloud and the IoT is depicted in Fig 1.

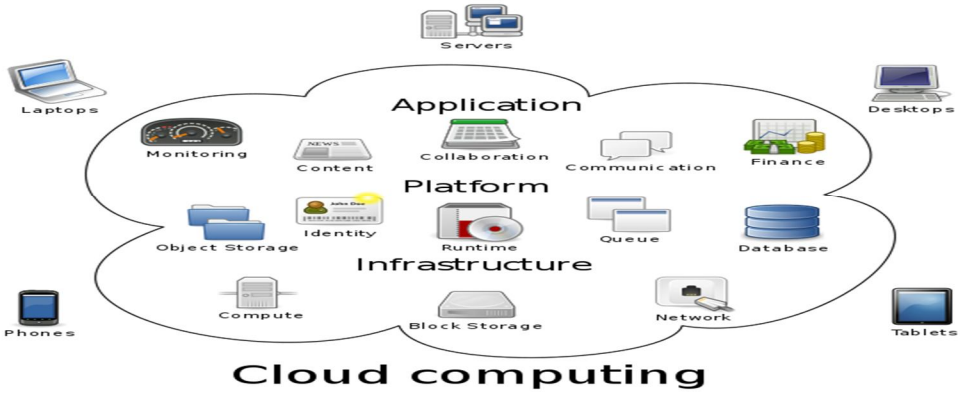

Fig 1. Cloud Computing
Source: https://en.wikipedia.org/wiki/File:Cloud_computing.svg#filelinks

However, here in this work the focus is on the storage part of Fig 1. The amount of data brought in by the IoT and the reason of the emergence of cloud computing and storage is attributed to the term called Big Data. As the name suggest it means very large

amounts of data. The IoT discussed previously dump a lot of data. With immense amount of data comes the opportunity to mine these massive amounts of data. But with this opportunity comes the potential of problems. There are a host of issues pertaining to Big Data which can be categorized into six namely applied ontology, security, storage and transport, associability, Inconsistencies, and Mobility.[6] But here the focus is on security. Security is a universal domain which spreads its tentacles to almost all spheres of information technology and Big Data is no exception. Data in the field of data mining is used primarily for unearthing trends and patterns in the given data which is called knowledge. The more the data the more precise the trends and patterns will be. Big Data eventually helps in this. Log reports collected can tell about the perform of a software or a sensor data from a airplane can tell about the problems in the aircraft in real time or a series of medical report can unearth patterns of symptoms previously unknown and which can potentially safe lives by early diagnosis and the list is endless.

Basically with Big Data any domain which has huge amount of data can harness those information for its growth. The flip side is that it presents and easy target for hackers who, if they can break into the storage of the cloud can take any unimaginable amounts of data, which in traditional methods would be time consuming. However these issues have now being addressed with strong cryptographic permutations and other safe guards. However security is an evolving field and a myriad of techniques and rules are used to enforce security. With cloud storage and the present modern day internet applications there is another security threat called privacy. Security ensures who can access your data whereas privacy means what one does with your data. Privacy is a deep rooted problem. For example a seemingly harmless looking photo of a person in a holiday can give a wealth of information which can be obtained from other objects in the background, like vehicle registration can tell where you are or a prominent landmark and tell where you had been etc. These are information that should be private and photograph taken wasn't indented to reveal these information's. When one comes into the possession of these information's, what they do is what determines privacy intrusion. It's a potent problem in today's world with the age of social media with deadly replications. [7-9] For security in Big Data there a lot of techniques like cryptography, association rules, auditing and various other techniques to enforce it. But privacy is a tricky proposition. A computer can't tell nor alter the mind of an individual upon receiving the information. What you can do is to make the information seem meaningless. For example take for example a medical record of a person as shown in fig 2. Only a tuple has been displayed for representation purpose. Take Fig 2 (a) and you can see the information of a heart diagnosis which doesn't yield any other information. Now take fig 2(b) and the zip code tells where this person is from. Fig 2(c) tells the gender and fig (d) tells the age, which when coupled with gender you can classify the person into young and old. The final field, name shown in fig (e) ties all the information shown in fig 2 (a-d) as one complete package. With the name you can tie this someone. With the others for eg age and gender you are able to tie this to a particular group and not an individual. Hence here a novel approach is been proposed to solve this particular problem by making information seem meaningless even though they are grouped together.

Diagnosis
Heart disease

(a)

Zip	Diagnosis
12201	Heart disease

(b)

Sex	Zip	Diagnosis
Female	12201	Heart disease

(c)

Age	Sex	Zip	Diagnosis
24	Female	12201	Heart disease

(d)

Name	Age	Sex	Zip	Diagnosis
Andy	24	Female	12201	Heart disease

(e)

Fig 2 (a-e) progression of meaningfulness of data

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

II. EXISTING APPROACHES

There are various approaches to solve these problems in literature, some of which are discussed below. K-Anonymity is technique where a data present in a table satisfies the K-Anonymity property where similar data are grouped in form of ranges that handling the age. For eg:- in the previous example stated above instead of represented age as a number it can be presented as a range.[10]

The i-Diversity concept attempts to prevent homogenous attacks. It determines how diverse the values are from the point of sensitivity. It is based on the entropy and skewness of the sensitive information.

The t-closeness technique aims to guard against a specific kind of privacy constraint by splitting the sensitive data into a set of equivalence classes. These spitted are spread across the table. [11-12]

The sample three approaches discussed above aim to split or decouple the data so that it stays anonymous. Privacy is attained when vital information also called sensitive information are kept away from the unknown. There are other techniques in the literature which all work in a similar fashion. Keeping in mind the work done previously here in this work a unique technique that splits the vital part of the field and makes it appear meaningless is proposed.

III. PROPOSED APPROACH

In the example given in fig 2 it is imperative that the name field is the one that binds all the other field together. Without the name you would looking at a broad range of data rather than tying it with one person. In any database you look there would be one field that would enable you to tie the whole information to an individual. Take for example mark statement. When the name is attached with a series of marks and other information it reveals more about the individual. If you avoid name and put reg no instead then no one know mostly to whom it refers to. The same sort of analogy is used here.

The proposed approach is a two level systematic approach. It aims at splitting the meaningful field in a tuple into two parts to make the field meaningless. The processed approach is show in Fig 3.

Often the data or dataset is fetched and the data as we know are arranged in Tuples and further divided into fields. Even when a dataset is requested the back processing involves retrieving the dataset by the data manager. Here in this approach T represents the tuple and T(f) represents the various fields in the tuples. Take the sample Tuples shown in Table 1. The Priority Level (PL) are designated on how well the particular field can relate to an individual. Name can often directly link to a person so its always high priority. Though the name itself cannot reveal much information, name tied with any of the other fields can reveals a lot. For example name tied with age reveals that it is a young person, likewise. Most of the techniques try to frame rules and enforce cryptography techniques but here a unique hashing method is employed, which splits the name into two words making it eventually meaningless to the onlooker.

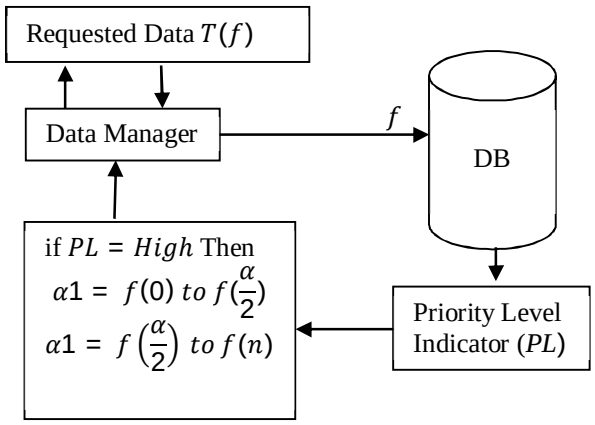

Fig 3 The working of the proposed approach

Table 1 A sample medical dataset

Name	Age	Sex	Zip	Diagnosis
Andy	24	Female	12201	Heart disease
Tom	23	Male	14302	Diabetes
David	26	Male	18310	Kidney disease
Nash	29	Male	19522	Hepatitis
Alice	32	Female	22505	Diabetes
Mark	49	Male	24800	Hepatitis

The data requested by a application or anyother source either genuine of non-genuine sends a requests for a set of data which is arranged in the form of tuples (T) in the database which is divided into fields (f). The data manager retrieves (f) from the database for every request of T(f). Depending on the sensitivity of the data the priority levels (PL) is marked. In any dataset as discussed earlier name gets the highest priority. Then the highest priority data which is the one that helps to pinpoint T(f) to an individual is spilt into two based on the mid square hashing method.
The idea is that for eg:- a name sandy which sounds like a name will not sound like a name it if is san or dy. It is much like gmail password checker which looks for meaningful names. When the technique is on table one we get the result as in table 2

Table 2

Name1	Name2	Age	Sex	Zip	Diagnosis
An	dy	24	Female	12201	Heart disease
To	m	23	Male	14302	Diabetes
Dav	id	26	Male	18310	Kidney disease
Na	sh	29	Male	19522	Hepatitis
Ali	ce	32	Female	22505	Diabetes
Ma	rk	49	Male	24800	Hepatitis

As you can see the name appears meaningless as shown in table 2. You could program an application that wants to use this data to use the two fields to create name if it needs. Otherwise to an ominous onlookers it dosnt give any meaning at all.
The priority level (PL) is determined on the basics of sensitivity which means how well a particular field when coupled with other fields changes its potential. The results are shown in Table 2. Here two priority levels are determined namely high and low and they are manually determined.

Table 3 PL levels for table 1

Name of the field	Priority level
Name	HIGH
Age	LOW
Sex	LOW
Zip	LOW
Diagnosis	LOW

Now in table 3 the High priority is assigned to Field 1 which is name. The meaningfulness if these names are show in Table 4. Based on the meaningfulness of a name a value of 1 is assigned if it is meaningful and 0 if it not and the percentage is calculated. Table 5 shows the same after the proposed approach is implemented.

Table 4 Meaningfulness of names in table 1

Name	Meaningfulness
Andy	1
Tom	1
David	1
Nash	1
Alice	1
Mark	1
% of Meaningfulness	100%

Table 5 Meaningfulness of names in table 2

Name	Meaningfulness
An	1
To	0
Dav	1
Na	0
Ali	1
Ma	0
dy	0
m	0
id	0
sh	0
ce	0
rk	0
% of Meaningfulness	30%

As you could see the key to privacy is making the data seem meaningless so that the onlooker cannot tie the information together. Here if you compare the results obtained in table 3 and table 4 we obtain a reduction of 70% in meaningfulness and it is depicted in graphical format in fig 4.

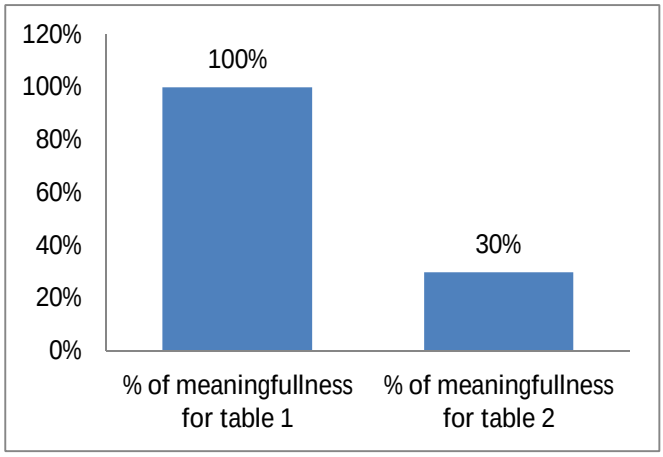

Fig 4 Comparison of Meaningfulness

IV. CONCLUSION

Here in this approach a novel approach is being tried and tested that ensures privacy by making the important data meaningless. Privacy is a major challenge as you can't stop nor predict what one would do with sensitive data. By this approach you can be sure that no one uses sensitive information as the data presented is made in informatic. In future more such methods could be evolved to preserve privacy and a combination with security would provide a worthwhile system.

REFERENCES

[1] Hassan, Qusay (2011). "Demystifying Cloud Computing" (PDF). The Journal of Defense Software Engineering. CrossTalk. 2011 (Jan/Feb): 16–21. Retrieved 11 December 2008.
[2] Peter Mell and Timothy Grance (September 2011). The NIST Definition of Cloud Computing (Technical report). National Institute of Standards and Technology: U.S. Department of Commerce. doi:10.6028/NIST.SP.800-145. Special publication 800-145.
[3] M. Haghighat, S. Zonouz, & M. Abdel-Mottaleb (2015). CloudID: Trustworthy Cloud-based and Cross-Enterprise Biometric Identification. Expert Systems with Applications, 42(21), 7905–7916.
[4] Santucci, Gérald. "The Internet of Things: Between the Revolution of the Internet and the Metamorphosis of Objects" (PDF). European Commission Community Research and Development Information Service. Retrieved 23 October 2016.
[5] Inukollu, Venkata Narasimha, Sailaja Arsi, and Srinivasa Rao Ravuri. "Security issues associated with big data in cloud computing." *International Journal of Network Security & Its Applications* 6.3 (2014): 45.
[6] Samuel, S. Justin, et al. "A survey on big data and its research challenges." *ARPN Journal of Engineering and Applied Sciences* 10.8 (2015): 3343-3347.
[7] Matturdi, Bardi, et al. "Big Data security and privacy: A review." *China Communications* 11.14 (2014): 135-145.
[8] Thuraisingham, Bhavani. "Big data security and privacy." *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*. ACM, 2015.
[9] Bertino, Elisa, and Elena Ferrari. "Big data security and privacy." *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*. Springer International Publishing, 2018. 425-439.
[10] Sweeney, Latanya. "Achieving k-anonymity privacy protection using generalization and suppression." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 571-588.
[11] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.
[12] Davis, John S., and Osonde A. Osoba. "Privacy Preservation in the Age of Big Data." (2016).