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

An Effective Framework to Provide Optimized Security in Big Data Environments

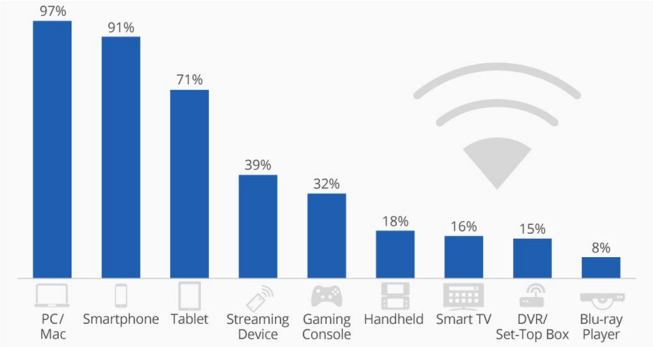
D. Angelin Benita[1], S.P. Victor[2]
[1]*Research Scholar, Department of Computer Science, St. Xavier's College, Palayamkottai, India*
[2]*Department of Computer Science, St. Xavier's College, Palayamkottai, India*

Abstract*: *The Buzz word in the industry of late is the Big Data. It is the most talked about term across multiple domains which directly or indirectly profit from the usage of this technology. In short Big Data can be called as an enlarged version of data mining. It can be said as such because the size quantum of data being used here are exponentially huge. Not only is the size of data is huge, also the problems we encounter in data mining also magnify into huge proportions and the worst news is that the traditional methods used in data mining won't work for Big Data, simple because the size of the data is huge. Hence it throws open a huge field with varied research potential. Here in this work, a novel approach is presented to tackle the problems pertaining to security. Security implementation methods are a double edged sword. Too much security leads to inefficiency and too little leads to compromise of data. Balancing this is the key to success in any application. But in Big Data apart from this balancing act, the huge quantum of data makes traditional approaches of security time consuming. Hence a unique strategy is proposed here that not only provides the right security but it can also deal with the huge quantum of data.
Keywords*: *Big Data, Security, Internet of Things.

I. INTRODUCTION

Big Data is a term that defines data sets that are exponentially large and complex, so much so that the methods that work well with traditional data are inadequate to deal with this kind of data quantum's. Of late the term "Big Data" has been associated with predictive analyzes and user behavior patterns. The challenges presented by the Big Data range from data capture, analyzing data, searching, transferring, viewing the data, querying and information security and privacy. The applications of Big data range from prevention of diseases to combating crime to spotting of trends in business etc[1].

The growth of Data Sets is rapid these days, the reason for it being the contribution of data by the so called Internet of Things such as sensors, smart phones and other mobile devices, software and network logs, cameras, cameras etc., Any device that can be hooked up to the internet joins the family of the Internet of Things. Fig 1 shows the contribution of data obtained from various devices in the US for a sample set. The graph shows the broader picture of the various devices that contribute data. More data means better analyzes and better prediction. So in future the numbers of the so called Internet of Things are only going to increase. [2]



Source: https://infographic.statista.com/normal/chartoftheday_5172_household_penetration_of_connected_devices_n.jpg

Fig 1 Percentage of Contribution of data by IoT

The research potentials in the field of Big Data are divided into six areas namely applied ontology, security, storage and transport, accessibility, inconsistencies and mobility [3]. These stem out of the necessity in adhering to the characteristics of Big Data, which is depicted in Fig 2.

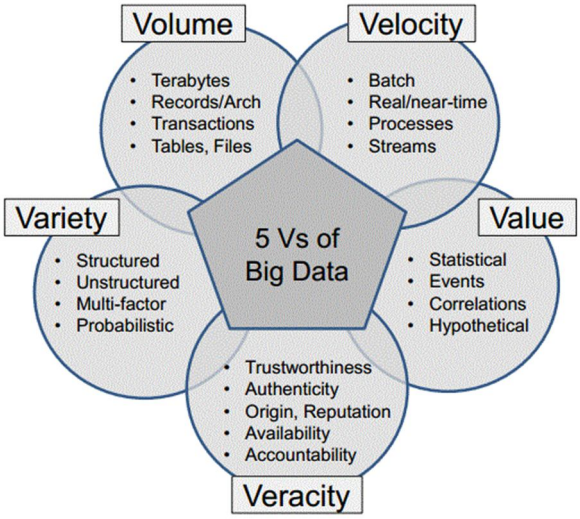


Fig 2 Characteristics of Big Data

Volume refers to the stored data after generation. The size of the data determines its potential. The more data you have, the better potential you have to do good analysis. Variety refers to the nature or better said as the type of data. Velocity refers to the speed at which the data is generated. Veracity refers to the authenticity of the data and value and Value is the worth of the data being analyzed. Together these characteristics give Big Data its shine. To achieve this various facets have to be polished hence the above mentioned research areas are crucial. In this work the focus is on security. Security can be described as almost a fence which restricts access of whatever content is there within it, to a few individuals with certain parameters. Security is a big deal in computers. Before networking became prevalent security wasn't a big deal. Because at most the only threat of your data being accessed is that if someone can physically lay their hands on it. With the advent of networking and internet the data in ones computer can go to places where one does not want to or be accessed by who should not. Since then the concept of security has taken a significant place. It has secured a place in the evolution of computer in making the internet and computers a much safer proposition. But time and again security measures can become nullified because of two primary reasons. One is that the technology or the field has evolved such that the security measure previously employed cannot cope up or that hackers have being able to nullify and breach the security measures put in place. In either case it is prevalent that security has to evolve with time and should be modified from time to time to suit the latest technologies, systems and domains.

Here the focus is on security in Big Data. When anything is stored in large volumes, it easily becomes a target for malicious elements. Because if they can nullify the security measures put in place then obliviously they can get access to huge quanta of data easily. You might consider keeping the data in separate location so that an intruder does not get access to the data on the whole. But the overhead involved in getting the data to the respective clients presents a huge challenge and this reduces efficiency. So from the above, it can be summarized that the huge challenge in incorporating security features is maintaining time and efficiency in transacting the data safely. Hence in this work, a novel approach is being proposed so as to maintain security in Big Data with compromising security and efficiency.

II. EXISTING SECURITY APPROACHES

The security issues of Big Data are broad. It affects almost every level in the data storage and retrieval process. Some of the key security issues and how they can be solved are discussed below.

In a distributed programming setup parallelism in both computation and storage leads to massive build up of data. In scenarios like this, algorithms like map reduce are implemented. So here there are two challenges namely securing the manner in which data are divided into chunks and also monitoring where they are stored. Big Data processing software are unable to perform these operations hence these have to be manually performed.

Since data is being manipulated, split and stored at different location the logical thing to do is to monitor the end points. This means the device providing data and accessing the data should be secured at the organization.

In Big Data, data are moved and dumped almost dynamically. In such a scenario keeping track of the various data is done by

maintaining transaction logs. Proper care has to exercise so that the data doesn't fall into hands of malicious elements.

In Big Data the data is being processed in real time and as such it is not feasible to maintain regular checks. Hence it's advisable that security checks be conducted as and when the data arrive.

Secured data storage is a great solution but often data storage devices are vulnerable. Therefore it is necessary to encrypt the access control methods as well.

To know if the given data is harmful or if would be a potential target for hackers, one has to be aware of the origins of the data, the authentication used, the validation process and the access control parameters and methods.

One of the most trusted security features spanning different domains is auditing. Auditing helps to keep a check and hence indicate if something is falling out of line. Here in the computer world logs are used to find out which data is being accessed by whom and other related parameters are monitored. Hence regular monitoring can thwart cyber attack or malicious activity quickly. Granular access control of Big data stores provided by Hadoop and NoSql requires compulsory access control and a strong authentication process. [4][5][6]

When we talk of encryption what comes to one's mind, is cryptography. There are proven solid techniques that provide great security to ones data. The most commonly used techniques are classified into two namely symmetric and asymmetric Symmetric algorithms use a single key for both encryption and decryption. The used encrypts the data using a single key and sends it and the receiver utilizes the same key to decrypt and retrieve the information. DES, TRIPLE DES, RC4, RC6, BLOWFISH, AES are some commonly used symmetric algorithms. [7]

Asymmetric algorithms on the other hand make use of two keys namely private and public key. One key is used to encrypt the data whereas a different key is used to decrypt the data. The private key is kept secret whereas the public key is distributed. The RSA and DSA are some popular asymmetric algorithms used. [7]

From the above survey of the literature the following things can be summarized.

1) Security of a data depends on the data being stored and to whom it belongs to.
2) Security is not a one stop solution, its needs a structured or layered approach.
3) A security approach should be able to adapt to varying needs
4) Big Data is voluminous and hence traditional security methods don't work.

To counter this problem here in this work a unique framework is proposed which would take care of all the issues summarized above and work in a way that it takes the best of the traditional methods available and make it work with the Big Data Environment.

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
30	unemploy	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	managem	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	managem	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-colla	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	managem	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
36	self-empl	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
39	technicia	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
41	entrepren	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
36	technicia	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
31	blue-colla	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no
40	managem	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
56	technicia	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no
25	blue-colla	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no
38	managem	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0	unknown	no
42	managem	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0	unknown	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	-1	0	unknown	no
44	entrepren	married	secondary	no	93	no	no	cellular	7	jul	125	2	-1	0	unknown	no

Fig 3 Snapshot of the Banking Dataset used here

III. PROPOSED APPROACH

The proposed approach works around the weakness posed by enormous quantum's of data by categorizing data into priority levels. The higher the priority the more critical the data is. Here for illustration purpose a sample banking database has been taken [8] as shown in Fig 3. Higher levels of security are needed for passwords and certain details like name and address need differing levels of security. Security is a double edged sword. One needs best security without compromising efficiency. Normally the tendency is to employ just one security approach to maximize efficiency. But not all fields of data need the same level of security. Here the aim to categorize data into different priority levels and implement various levels of security. As stated in the previous example of a bank, not all the fields are going to accessed at one go. First when the user enters the username and password are the ones needed and then upon the ones need the other information are accessed. The way the proposed approach works in show in Fig 4.

Here the categorization of is based on the how directly a field can tie the entire set with the individual. If a field T can directly identify the individual, like a name of phone number it is classified with Label H, meaning high priority and in turn the best algorithm is to be applied here. Anything that can indirectly refer a field T to a individual like for example age which tells which class that person belongs to or a set of tuples $F(T1, T2, \ldots TN)$ which can inference to a person are marked with label M, which states medium priority. And the final label is L which means low priority for which security need now be applied. The priority levels for the banking dataset is show in table 1

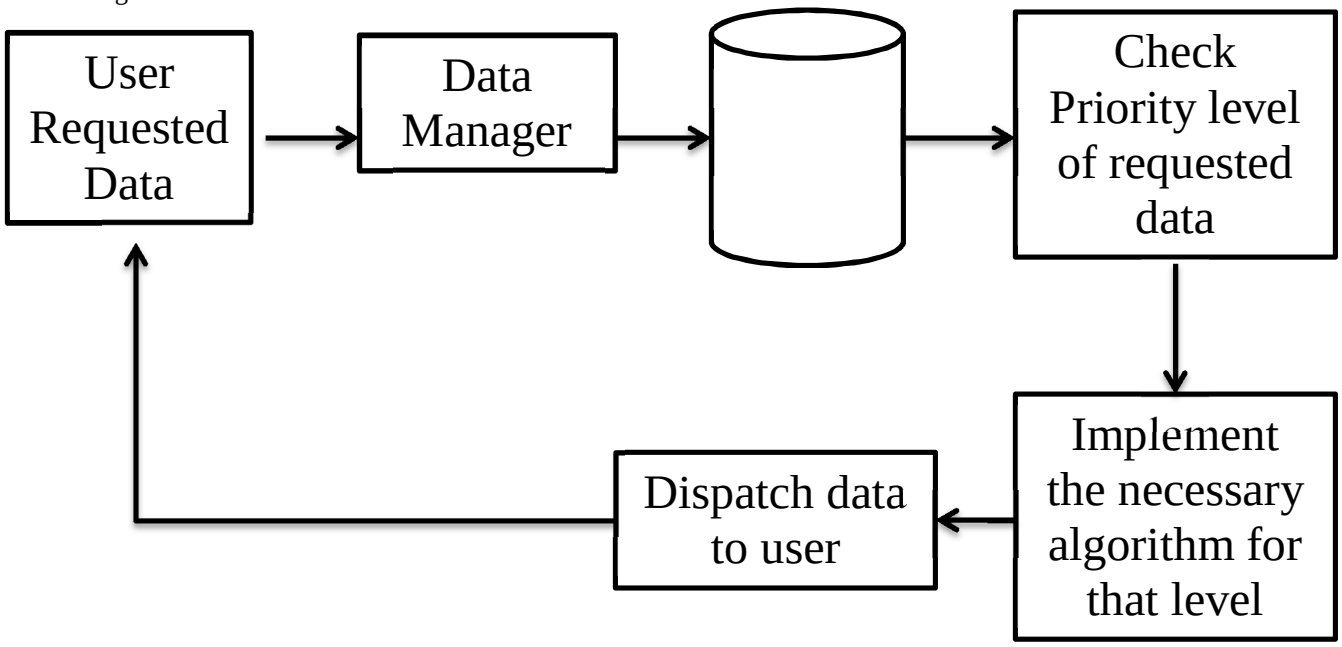


Fig 4 Block Diagram of the proposed approach

Symmetric algorithms as discussed in the literature take less time and Asymmetric algorithms take some extra time.

Once the priority levels are sorted out, its just a matter of preference as to what the end level application is out to do. For example if the main criteria of a security based algorithm is speed then you can opt for Triple DES, RSA, Blowfish, Two Fish, AES. If you are on the look out for the most secure cryptographic algorithms then you may well as go with AES, 3DES, Two Fish or RSA. If you want both speed and accuracy you might adopt RSA and Two Fish. It all boils done the application in use.

Table 1 Priority Level of Various fields

Field	Priority Level
age	H
job	H
marital	L
education	L
default	L
balance	H

housing	M
loan	M
contact	H
day	M
month	M
duration	M
campaign	L
pdays	L
previous	L
poutcome	L
y	L

IV. CONCLUSION

Thus here in this work, an effective frame work is being proposed to solve the problem of providing high security to sensitive datasets, by making use of existing proven techniques with less time by categorizing the dataset into distinct priority levels. The idea of using a framework helps us to incorporate various techniques that are good in other places, by tying them in a way that works well for our problem. In future more such frameworks can be evolved.

REFERENCES

[1] Boyd, D., & Crawford, K. (2011, September). Six provocations for big data. In A decade in internet time: Symposium on the dynamics of the internet and society (Vol. 21). Oxford: Oxford Internet Institute.
[2] Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.
[3] Samuel, S. J., RVP, K., Sashidhar, K., & Bharathi, C. R. (2015). A survey on big data and its research challenges. *ARPN Journal of Engineering and Applied Sciences*, *10*(8), 3343-3347.
[4] 10 Challenges to Big Data Security and Privacy, http://dataconomy.com/2017/07/10-challenges-big-data-security-privacy/ accessed on 19-10-2017
[5] Upadhyay, Govind Murari, and Harsh Arora. "Vulnerabilities of Data Storage Security in Big Data." *IITM Journal of Management and IT* 7.1 (2016): 37-41.
[6] Matturdi, Bardi, et al. "Big Data security and privacy: A review." China Communications 11.14 (2014): 135-145.
[7] Er, Manpreet Kaur, and Jasjeet Kaur Er. "Data Encryption Using Different Techniques: A Review." International Journal of Advanced Research in Computer Science 8.4 (2017).
[8] https://archive.ics.uci.edu/ml/datasets/bank+marketing#