ISSN No:-2456-2165

Data Governance Model for Tagging Data Using Finger Printing

Deepa Mahadev[1]
Assistant Professor, Dept. of CSE
BNM Institute of Technology
Bangalore, India

Asha K [2]
Assistant Professor, Dept. of CSE
BNM Institute of Technology
Bangalore, India

Abstract:- Data is considered as a raw entity which has impacted multidisciplinary research endeavors as well as government in developed and developing countries. Data is getting generated from a giant stock market to a tiny smart mobile phone. Hence, we are in the era of big data where data sets are characterized as Voluminous, variety, veracity, and Velocity.

Today generating data from various sources is a matter of time and requirement. The data collected are easily analyzed through various intelligent models proposed by researchers based on the application requirements.

Once data gets accumulated, one must take care of data governance, data ownership and data anonymity. With the advent in technology and growth of AI & ML, researchers did not concentrate on give back the analysis report to source of data generation point. Hence data governance, data ownership and data anonymity are all at their infancy.

To this end, the paper aims to develop a data governance model which takes care of data ownership and tags data with source of availability. This is possible if data is finger-printed at the source of generation, thereby the ownership is confirmed, anonymity does not exist, and analysis report is back with the owner for improvement or modifying the existing system.

Keywords:- *Data Governance, Data Ownership, Data Anonimity, Privacy, Data Finger Printing.*

I. INTRODUCTION

Big data challenges are multi-dimensional and are known by three defining properties which is so-called '3 Vs' (Data Velocity, Data Variety, Data Volume), which makes it more challenging to extract information and business analysis. In addition to those three features, when big data is processed, analyzed, and stored, other features like data governance, data ownership, data accessibility, privacy, and security policies (as shown in Fig. 1) comes into play. At present there are various definitions and important concepts to be considered for data and information governance. Efficient management of big data requires expertise and methodologies varies from those needed to manage "normal" data, which also needs a change in analysis procedure.

Companies recognize that there are varies challenges that they will face during data collection and as applications are operationalized and the issues being varied for different organizations, the issues are not the same for all organizations.

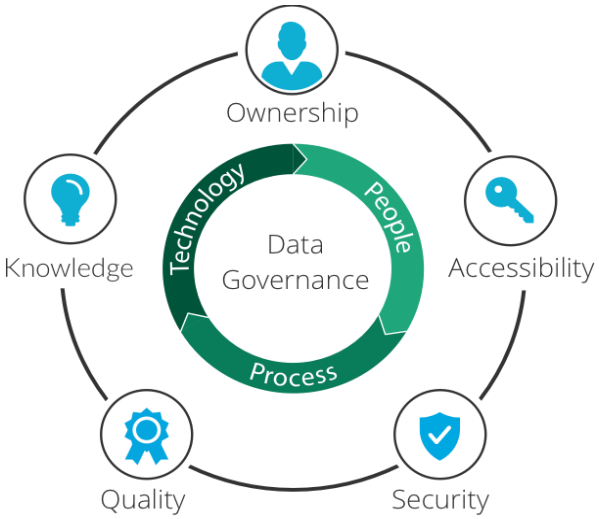

Fig.1: Data Governance and its challenges

At present, data collection, data access, data process, and data protection are all major concerns in the context of big data. These issues majorly include handling of personal and business data, availability, access and re-use of data, data security and privacy (Eg. timely updates), authorization of owners, cybercrime, approval of e-documents, liability for biased information, standardization of processes and policies.

Upgraded methodologies and processing power have given organizations the capability to build more futuristic and in-depth personal profiles based on one's online and offline detectable patterns. The data thus raised from such systems are highly monitored, recorded, and stored in different forms, for ensuring flawless user experience.

Thus, it is a requirement for companies to enhance the scope of current procedures for data governance to include big data and to develop a new data governance model if one does not exist, to support existing programs and be compatible with the applications. With this regard, it is thus required to develop a data governance model that takes care data ownership, anonymity and tags the data with the source of availability. This tagging of data using finger-printing method at the source of data generation, is a thought which is at its infancy.

ISSN No:-2456-2165

II. DATA OWNERSHIP

One of the big challenges in big data is to decide who is the true owner of the data? The common belief in many companies and public is that IT owns the data and people are just users of the data. But to a certain extent the belief is not true. The degree of data ownership actually depends on the level of useful analysis made from the data. There is much about what and how much part of data is considered as 'information' really in today's world. Data has intrinsic values along with added value, what one uses to provide some information. The context and the utility provide a meaning to the data that frames information. IT has more influence on the data; Companies may have their own control for developing certain methodology, but data definition, data production and data usage are not part of Information Technology supervision.

Though data storage to some extent takes place in IT, cloud storage, depending on different case studies and application, there raises issues who will be the owner of the data. Also, after data anonymization, as data content changes (after analysis) and integration of data also happens, it's difficult to give back the data and analysis and to traceback the owner of individual data. Thus, data in the isolated mode have no relevance and therefore no value. When there is no value in data, then one would deduce the true owner of the data and ownership is an issue. This results as the major conflict of data ownership.

III. DATA ANONYMITY

Personal privacy preservation is known to be one of the main challenges in the world of Big Data. Reason being the possibilities of revealing in sensitive or personally recognizable information while handling huge voluminous data. Individuals' Data protection and privacy preservation can balance with the challenges generated by the Big Data storage and analytics processing, with focus on anonymity is disclosed. In fact, even when dealing with anonymized raw data, sensitive information can be drawn out through analytics. Preserving anonymity is distinctly difficult as it should be done while allowing the analytics to produce meaningful and useful perception about the data.

Practically, the major challenge lies in the back-and-forth between keeping the data useful for organization while maintaining privacy and returning the data back to owner. The usage of data decreases, as the privacy requirement increases. Perfect usage can be obtained by publishing the data as exactly as received from data owner, but this offers no privacy; Perfect privacy can be achieved by publishing nothing at all, but this has zero usage.

Data Anonymization has procedures that can be applied to prevent the revealing of personal data. The Anonymization procedure is carried out to change the data before its being disclosed. This is majorly needed to reduce personal information leakage of an individual while data being communicated and shared to Public.

IV. CHALLENGES IN DATA ANONYMITY AND DATA OWNERSHIP

a. Anonymization Policy standard Definition.

Like privacy policies, anonymization policies surround what type of data fields must be anonymized. The major challenge is the need of anonymization policies to be followed and to find out better ways to protect the privacy of persons and their data.

b. Efficient Data Anonymization Methods

Each data analytics applications' have their own goals. In many situations, it is important to select the more adequate method to be used according to the context of application, even though there are different methods of anonymization. Along with raw data anonymization, it is also required to measure and govern the analytical disclosure of the data analytics algorithms. In fact,

c. Risk measure of data usage and publishing

A major issue for de-identification is to ensure protection of data with minimum loss of accuracy. Practically, the more the data is anonymized, lesser it is useful to the recipients as few analyses becomes impossible, or the analysis produces wrong and confused results. Thus, it necessary to measure the data usage of an anonymized database; the information extracted from an anonymized database must remain relevant and intact. One other important measure is the publishing risk. There are chances of data disclosure, i.e., data de-identification, even after anonymization techniques are applied.

Thus, in all these voluminous generation of Data and delegation of ownership for the data that is being generated, where actually does the ownership exist? A new set of data related to the information that was generated, will be created every time. Further dependency on the data, will raise in thorough scrutiny and verification. The idea here is that the source which can verify the data generated and confirm the correctness of the information will be considered the 'True Owner' of the data.

But, after the process of Data Anonymity, the set of question rises is to how to identify back the "True Owner" of the data. This raises up the critical challenge of data utility and information utility.

V. DATA FINGER PRINTING AT THE SOURCE OF GENERATION

Data Fingerprinting may be used as novel dimension to balance between privacy concerns and public disclosure of data. So, in this regard, there is a necessity to develop a Data Governance model which takes care of data ownership and tags data with source of availability. This is possible if data is finger-printed at the source of generation, thereby the ownership is confirmed, anonymity does not exist, and analysis report is back with the owner for improvement or modifying the existing system.

ISSN No:-2456-2165

Data can be represented in a unary style; the identity of the bits can be used as a printing pad to blemish the data with the handler's identification. Data which is Passed from handler to handler, identifies its route, and authenticity. Data which has voluminously traveled on networks will collect the 'fingerprints' of its handlers, to allow for a forensic analysis of fraudulent attempts, or data misuse.

The basic idea is that all data can be expressed as integers, all integers can be represented as a series of bits where the count of bits reflects the data carried by the string. Accordingly, all 2^n possible n bits strings will carry the same value, n. The range of 2^n possible strings all representing the same value n, may be used as meta data associated with the prime data (n), and this meta data may be regarded as 'fingerprinting' the primary data, n.

The meta data which is considered as "fingerprint" for the data is the personal identification of the "True Owner" of the data. This meta data is considered as "Tag" for the data that travels through out the network. Thus The "True owner" of the data remains to be the source of the data. Also, even after data anonymity, data analysis the ownership of the data will not be lost because of the availability of the "Tag" throughout its pathway.

Two major methodologies can be used to connect any single term to a row of data. Usage of regular expressions being the first one, works well for something like smart card numbers, where number of digits is expected more and as digits have a consistent set of repetitive data. "Value-based" fingerprinting being the second one takes basic personal information about the row, the data type, the value distributions, etc. in a specific permutation, and also a numerical formula can be framed for converting the basic unique information into a unique identifier, which is referred to as a "owner signature." Then comparison between identifiers can be made and find other rows that match the identifier. The simple logic here is to get the algorithm for generating the unique pattern to be considered as "Owner Signature": the right signature just to avoid both biased and incorrect matching and tagging to authenticated owner of the data.

VI. CONCLUSION

The words like 'data ownership' and 'data anonymity' is likely to seek more attention from both practical and research fields of big data. The organizations will continue to use data as a source of competition and prosperity. The paper also discusses the major challenges regarding privacy and security in big data processing and its analysis. Addressing these referred issues is required, to make sure the responsible and continued privacy protection, supporting privacy research. With this regard, paper attempted to present a simple idea of tagging user data with finger printing that bit identities, rather than carry the primary data, will be used for tracking, security, and general forensics.

REFERENCES

[1]. Alstyne, M.V., Brynjolfsson, E. and Madnick, S.E. 1994. Why not One Big Database? Principles for Data Ownership.
[2]. Ballard, C., Compert, C., Jesionowski, T., Milman, I., Plants, B., Rosen, B., Smith, H.: Information Governance Principles and Practices for a Big Data Landscape. IBM (2014)
[3]. Ali A, Qadir J, ur Rasool R, Sathiaseelan A, Zwitter A, Crowcroft J. 2016. Big data for development: applications and techniques. Big Data Anal. 1(1):2
[4]. F. J. Kongso "Best Practices to Minimize Data Security Breaches for Increased Business Performance." Walden University Scholar Works.
[5]. of Information Privacy and Security Vol 13, 2017 issue 4 3. L. Cheng "Enterprise data breach: causes, challenges, prevention, and future directions"Wiley Online.