Collection and Usage of Data in Ethical Manners

Vasu Gupta
Dept. of Computer Science Engineering
SRM University, Delhi-NCR, Sonipat, Haryana
Delhi, India
vasuonemail@gmail.com

Nishant Sethi
Dept. of Computer Science Engineering
SRM University, Delhi-NCR, Sonipat, Haryana
Delhi, India
sethi.nishant43@gmail.com

Abstract—In the digital era today, data is the key for companies to understand their customers and their behaviors. But as per the power of data is understood, data is a valuable asset when it comes to interpretability of it into knowledge which could be used in any manner by the organizations for any purpose. The CUDEM refers to those problems and thus focuses on the need for defining the much needed ethical norms and boundaries to be defined over the retrieval/collection of data and it's usage. CUDEM method is a process of refining the process of acquiring the data of a person/customer by an organization/company up to an extent where it does not breach the privacy of the people/customers, and the data collected is sufficient to be inferred into their research database or the machine learning models for various commercial purpose usage of their data. Hence this method signifies for a solution that could be achieved when big data analysis is concerned keeping in mind the data privacy as well. As information is the new Nuclear power in this period, CUDEM is one way to keep it maintainable.

Keywords—CUDEM, Big Data, Big Data Analytics, Data privacy

I. Introduction

Data is something which can't be thought through, it is generated or produced every second and in a quantity which isn't quantifiable anymore. Fields like Big Data and Big Data Analytics came into picture when the pool of data became bigger and bigger and the need for the answer to the question raised -"What to do with it?"

But almost every organization or the company in the big data industry found a pretty much generalized answer to the particular question which is to be able to use it in any manner which benefits the commercial growth either in terms of profit (for profit-based) or improvement/upliftment (for non-profit based).Now this implemented the idea of another requirement which is to collect the data in order to be able to use it. So, the second question of the hour became "How to collect it?"

The answer to this question was found in form of creating/generating data-streams all across the internet platform where the activity on the world wide web by the individuals were being collected by the very browsing platforms they started /chose to access the web through.

Then the trend being followed in loop of collection and usage of data which in-terms became humungous to just let be stored in some databases and use, the improvised version of "usage" of this big amounts of data was introduced in the form of being able to sell it. The database became one of the most valuable set for a company who collected and updated their database for years because now the big data industry saw it as an asset of service which could be the next big sellable trade. This started the phenomenon for the new age activists to discuss "Is it really possible to trade someone's data?"

That question asked the worthiness of data , which at times made people dubious that is it really going to make their lives better or the model of selling data is just a fancy term which is no use for them and the only beneficiaries would be the big data concerning industries with their data heads analyzing the data of masses and what meaningful could be extracted out of that. But this phenomenon remain into a subtlety for a long period of time and under-rated as well but only until the scams unleashed the bitter truth of the data biasing for commercial growth and control popped out and all the eyes from the masses were answered the worth of their data as their identity which needs to be used in a certain mannerism for the right uses only.

Which brings us to the current question in this digital age-"Is our data collected and used in ethical manners?"

This evidenced-but-yet to be answered questioned is diplomatically answered by many professional heads of the related big data industries with their policies and spokesperson, to the world where a sense of doubt still remains.

Now, shifting the narrative to ethics.
For various industries, agencies and others whose role is to reach and involve the customers and the peers connected with them, access to the non-stopping and ever producing consumer data is vital. Among consumers, there is a growing expectation that corporations and brands will provide relevant offers and information, specified interests,

special attention and treatment— all of which are reliant on data. To assure their customers about their data, companies have to take on responsibility for making decisions about how they access and use data every day. Every profession relies on professional codes of conduct, the amount of things that any company can record about their users is substantial and using it in the most appropriate ways is the duty of the company's techs. For any breach of data or selling of data, the company personals are responsible and should be held accountable for[1].

Big Data within a larger system of firms, organizations, processes and norms for analysis. To create and use these large data sets to maximum effect, many firms aggregate data to create a new "whole" and sell access to this new data set. These so called the big data brands work cohesively with people agreements which is Big Data for customers—similar to any other industry. In this responsive phase of data heads such as CDOs which goes for Chief Data Officers are shifting to an outward, strategic focus in leveraging Big Data rather than the inward, service focus used for traditional data. Currently, however, there are not yet any industry norms or supply chain best practices that can guide them. By addressing the issues related with big data, the focus of this research paper also aggregates on the need to acknowledge them and hence root for the actual solution method which is efficient and follows the domain with an ethical approach. Hence, the data is the investment in a product—e.g., the software industry, the ERP industry, the automobile industry, etc. Importantly, if a market exists for a product, then a corresponding industry exists to meet that demand. Similarly, the market for Big Data is growing at an exponential pace and be measured, the corresponding Big Data Industry, comprised of those firms involved in the production, analysis and use of Big Data, begins to coalesce around standard industry practices[2].

As the data continues to stream in, companies should monitor and analyze the use of this data while considering ethical implications behind it, as well as privacy concerns. Looking into data ethically requires a systematic approach as well as a values-based analysis. Ensuring Ethical Use of Data with a set of rules to be followed. As with any rapid change, the growth of technological advances can not be matched with governments regulations and often this gap increases upto the level where nobody knows the implications. For example, consider the increased use of cryptocurrency. So far, not much regulation has been implemented to govern its use, which has left open the potential for illegal transactions to take place[3].

Therefore a need for businesses has also finally raised to ask themselves that will they be able to perform if the data usage or accessing benefits the consumer in such a manner that will also benefit in their perspective. Also they might understand this soon enough that mutual dependence on data is something which is going to dwell for both sides, either customer or company. Both the C's are benefitted[4].

Joining humans with A.I. might be the best way to make sure that ethical practices are not violated – AI can discern human error, while humans can use their critical thinking skills to evaluate each situation as it arises.The data analytics field is creative and is steadily growing, and it's an exciting time to enter the field with numerous possibilities. With an expertise in data analytics, it's possible to explore emerging career opportunities that can help shape industry change around the world[5].

II. Issues concerned with Data

- Privacy:- Privacy is something which is personal and has a limited bounds of sharing. Whereas the data part is concerned, the data privacy encompasses that someone's identity or their behavior is such a data which they have right to keep personal or private. The privacy is compromised if the data is analyzed for each individual to not only be able to recognize them but also analyze them and forming a judgement about their personality. If someone's privacy isn't respected, it becomes very close to exposing a person to a potential number of threats unknowingly.
- Security:- As far as security of data is concerned, databases are still vulnerable by potential hackers to be back-doored into and altering with the data stored in the databases of the system of the corporations, once a person achieves success in bypassing the security architecture of their systems. The security of the people or the diplomatic leads could be threatened by any rogue entity if it is compromised. This could also lead to data blackmailing.
- Risk of losing anonymity:- Platforms like deep web might sound much needed way of surfing the internet today but the surface layer web which is still used by the majority of the world can also, maintain that status at an extent if CUDEM is implemented and practiced.
- Lower rate of trust:- Data booms the confidence for any study or a research to reach for a conclusion which also indicates that something without any viable and previous dataset would be felt unappealing and less or may be unbelievable at time and this would set the basis of selection for anyone such as a candidate who is applying for a job or like a loan applicant who is applying for a loan. The greedy data feed will reduce the trust for an employer or the bank to give the job or loan respectively to the people who have less of data to be able to interpret about their preference and personalities or with an abundant amount of data in

form of a past record and current status of the person, people might not be given the choice or point to speak and convince and possibly a chance of say for a might-be-a-potential candidate or applicant being selected by the employer or bank respectively decreases. Which means dependency on data could increase more and trust on people could lose more.

- Loss of Uniqueness:- People with unique ideas and creative inputs help the society in many ways but what if the data collected for any unique idea or an invention becomes so redundant over the web that if a person with a unique ideation or an innovation claims it to be unique and might want to run for a patent filing , then it becomes entirely hectic and a time consuming task to research throughout the big data in the world where there isn't any possibility that the very idea or the invention had been already patented or in the procedure of getting one. Thus, large amounts of data is also a problem for the task of evaluation and it is unpredictable that the ratio of uniqueness will see an increase or an opposite of it could occur in the future.
- Corporate control:- As far as the modernity of this era is considered, it is hard to believe the corporate giants for trusting with something which belongs to self and only self. As the many existing harsh case studies conclude with the judgement about the big data industries tend to get sometimes greedy and careless regarding the upholding the trust of their very own customers who rely on them for their data privacy. The sneaky part is that the data is the power which the corporates use not only just to understand their customers but also to recognize their preferences, patterns and behaviors. This amount of knowledge is enough for them to interpret a profile for each of their customer and decide a recommendation system on the surface web to "control" what type of advertisements they see and how likely they should be approached so that they behave according to their expected graph of psychological behavior which was interpreted by capturing/spying or in the form of collecting their online activities(activities performed on the world wide web when surfing/interacting on internet).
- Violation of Rights:- Data collection and usage out of a certain bounds being defined can lead to violation of rights of the individuals as the Right to Privacy is every citizen's right and no one can or should violate it. It means, it should be a crime when this right is violated, this can bring serious damages to anyone, either being the direct target to be facing it or being the even indirectly, being the connection related with the one with an exposed identity one. This violation is damaging for a victim's relatives too.
- Spying Culture:- Trend of following and keeping a check on each other's activities online could often be adopted in future if the data is easily accessible to any kind of party.
- Data Royalties:- A phenomenon which might sound weird in present but may become the next big concern for everyone as the data royalties could be seen as a means for another type of business where people might forget the authenticity of their own data and sanction for willingly agreeing on selling of their data to third party companies in the luring model of being provided royalties for their data in exchange for its use in their commercial purposes.

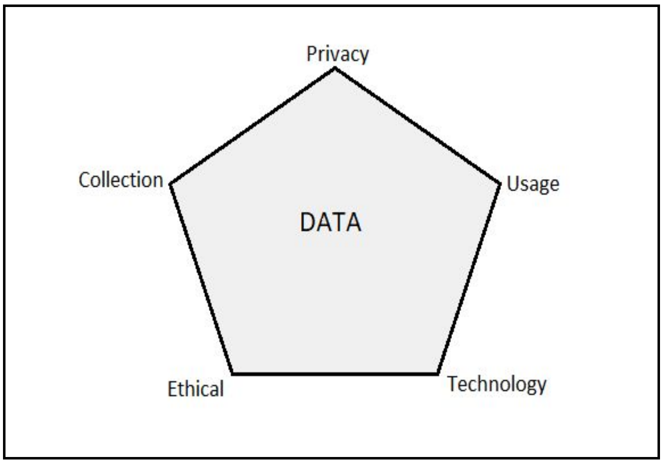

Fig. 1 *Correlation and bounds of data*

III. Current Industrial Techniques used in Data Collection

- Social network analysis
- Online content analysis
- Web-based experiments
- Online clinical trials
- Online qualitative research
- Cookies
- Cyber-ethnography
- Online interviews
- Location(GPS Tracking)
- Online focus groups
- Camera
- Online questionnaires
- Microphones
- License Plates
- Heatmaps
- IOT sensors
- Credit or Loyalty Cards
- Social-Media Activity
- In-Store Wifi Activities

With multiple ways to track any personal, There needs to be different boundaries that the corporates have to adhere to so as to have enough data for their needs as well as keeping user's privacy in check[6][7][8].

IV. Working of CUDEM

CUDEM(Collection and Usage of Data in Ethical Manners) can play an essential role in defining for a solution to this problem which could be implemented if the CSR(Corporate Social Responsibility) is seriously considered in terms of data as well. Data is needed to treated as ecosystem for a sustainable development for every sector. Thereby , CUDEM method is more applicable when like an eco-drive, corporates initiate volunteering for themselves in the pledge for collecting and using the data of people and their customers under some specific ethical bounds and norms.

Moreover, CUDEM method understands the need for both, that is, the need of collection of data and its usage for research purpose and also to be used for money based model for companies which are core in Big Data field to be able to be in the business and see their profits coming. Hence, scrutinizing this thin line difference for need of data to still be collected by these big industries for the necessity of this digital age to run as smooth as it runs today, this mutual relationship can't be denied that both sides need the data collection should still be running instead of just being stopped by, as undeniably, today, if we need to see progress, advancements and development in variety of fields, we will have to let the big data industries to collect the data which in turn will help make world a better place. This is why

CUDEM is the alternative method of approaching this thin line where both the circles are satisfied and a valid bound is also formed in which the system can work fine, and trust is able to be established neutrally without the fear of getting a data biasing situation. Also it is also necessary to work co-linearly where data biasing in discouraged and the data equality of treatment could be a sign of treaty for anyone with a sheer will and vision to use data in the right manner and thereby respecting the norms of managing it.

Moreover , CUDEM methodology is workable in almost every situation as it defines the definition of clean business without being greedy with such an abundant of amount of assets of data when available in the market. And the area of concern for the corporation which is profit can still be sustained.

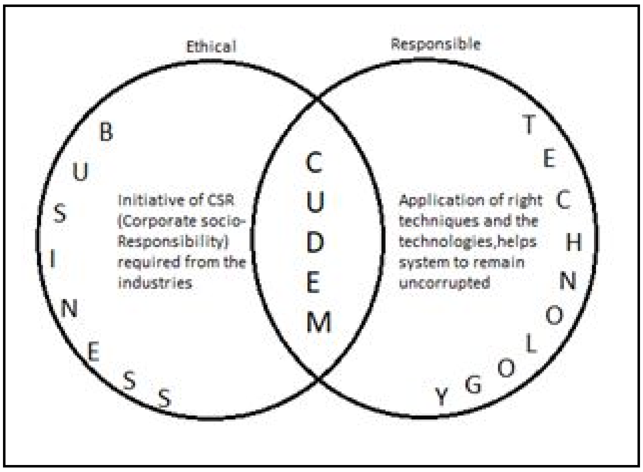
Ethical
Responsible
BUSINESS
TECHNOLOGY
CUDEM
Initiative of CSR (Corporate socio-Responsibility) required from the industries
Application of right techniques and the technologies,helps system to remain uncorrupted

Fig. 2 *CUDEM graph*

There is no denying of the fact that every company needs data. For some companies, data , is more or less the source of income. CUDEM is not asking a company to stop collecting any user data but rather, it is an approach that is making sure that the data collected is meaningful and at the same time is not infringing with privacy of any individual.

The key lies in unclustering of the personal identity in the so-called clustered data in machine learning terms. The machine learning models that are used in collection and clustering of similar type of datas combined to segregate into classes of hyperplane which classify the learned user behaviors and patterns.

Also with a user id which only a referred number to a person whose data will be tagged creates a layer of at least other third party companies to not be able to know our name, location etc., and the chance of identity leak could be saved. The "User Id" in the form of a DBMS(Database Management System) is going to be stated as an entity whose attributes are represented by the following: "Activites", "Profiles", "Preferences", "Location".

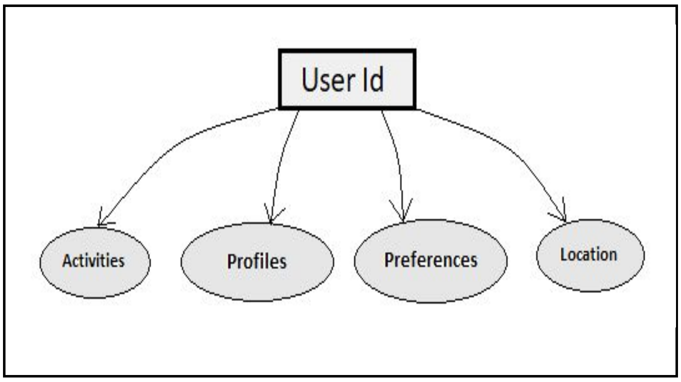

Fig. 3 *Entity-attribute relationship*

If the name of an individual is untagged from their specific dataset then the company only knows a user with so and so activities. Thus, social media profile linkages play an essential role in this phenomenon.

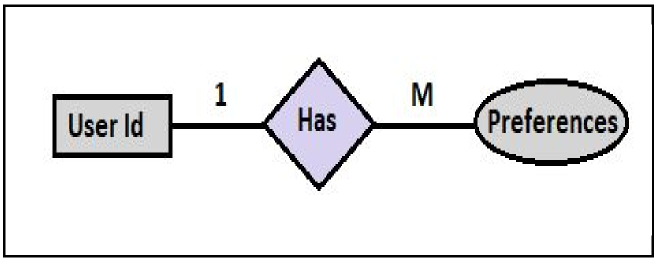

Fig. 4 *1:M (One-to-Many) Relationship*

Also the concept of entity-relationship encompasses the cardinal relationships that can be defined between an entity which has either one or many(more than one) attributes. These are known 1:M(one to many) relationships. Same can be understood by the above diagram. Fig. 4 shows that in a database model a User_Id can be relates/linked with numerous online activities.

Interlinking and exchanging of data in cross-platforms gives the harsh control to each company in very vast majority, meaning that neutrality of surfing the internet could be jeopardized by the companies and browsers by redirecting you or again and again feeding your eyes with the same kinds of advertisements or even control your psychology and might get you irritated by the recursive phenomenon of the same or the similar ads and sponsored posts. So interlinking or selling/trading of data with each other company shouldn't be practised as this not quite neutral for keeping the sanctity of the surface web and the data itself. Corporates fighting for data must generate their own streams for collecting it so that instead of many , only one knows the data and could use only in the ethical growth of their business and also help the customers.

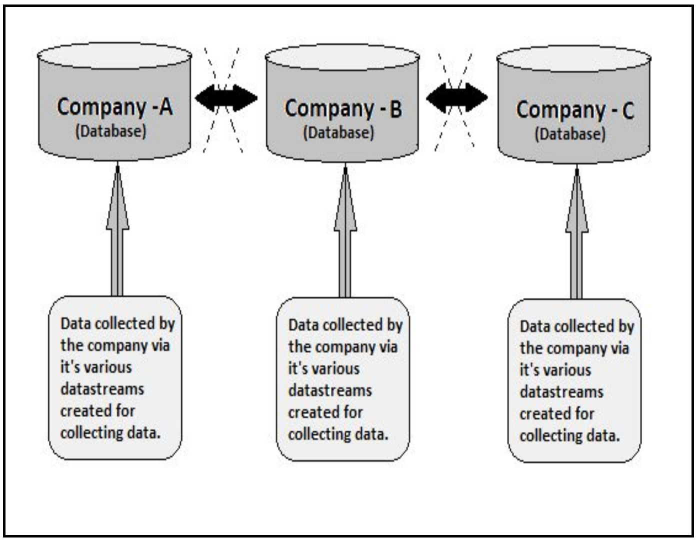
Company -A
(Database)
Company - B
(Database)
Company - C
(Database)
Data collected by the company via it's various datastreams created for collecting data.
Data collected by the company via it's various datastreams created for collecting data.
Data collected by the company via it's various datastreams created for collecting data.

Fig. 5 *Advocation of interlinking of data*

This can be done by setting different licenses/ rules with different kinds of data privileges. The company may choose the license which suits its requirements best.

The main idea is to not include any information of an individual that could track him/her in any way. This includes the personal information like name, age, address, mobile number. Instead of personal information, The model of a company can have a hash deployed to every user to see his/her usage pattern. What this means is that, Say for a Search Engine, if the company wants to know the user pattern, Instead of knowing the pattern of exact individual, they would have an estimation that people who search for query 1 and query 2 search for query 3.

These rules have to be defined by an expert panel which takes into consideration all the important data that is utterly required to be stored and the data that have to be neglected for example personal information.

CUDEM can not be forced onto any company rather it can be seen as a ethical mark that a company believes in.

CUDEM method may seem to only be user benefitting but it is not. If a users comes to know that the company is not storing their personal information for commercial purposes, the user will have much more confidence using the companies product.

Selling data shouldn't always result in hazardous and problematic situations if the data traded for the right purpose. This could be for calculating general facts and figures in order to compute generalized results by either the government or some non profit organization and NGOs. There are many future unprecedented event which we remain unaware of but their occurrence is guaranteed by some or non.

V. Conclusion

Henceforth, CUDEM method very much can be one of the future solution for the big data industry and the common people to be able to work in a trustworthy environment where data management becomes more and more safe and efficient and the right of privacy could again see to it's right light to the globe. Efforts from both side can make this system an appropriate cause of a "datanomical" future where issues concerned and related with data/big data could possibly and decrease and one could expect a data-driven world which becomes hopefully next to impossible to be able to be corrupted.End-to-end data sourcing can be made more safer and the control can be brought back to its very center and then the data neutrality can be expected out. This ethical manner of doing business with data will open the doors for unimaginable future possibilities of data-based businesses.

VI. References

[1] Jerry C. Jones, "Why marketers must shift the conversation from data privacy to ethical data use", March 1, 2018*(references)*

[2] Martin Kirsten. (2015). Ethical Issues in Big Data Industry. MIS Quarterly Executive.

[3] Jacksonville University, "How to Use Data in an Ethical Manner",www.jacksonvilleu.com/blog/business/ *(references)*

[4] http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf *(references)*

[5] Wanbil W. Lee,Wolfgang Zankl, Ph.D. and Henry Chang, "An Ethical Approach to Data Privacy Protection", ISACA Journal Volume 6, 2016

[6] The University of Minnesota, "Data Collection Techniques",https://cyfar.org/data-collection-techniques *(references)*

[7] Gaurav Jha, "4 Data Collection Techniques: Which One's Right for You?", August 16, 2017 *(references)*

[8] Melody Ucros, "10 Sneaky Ways Companies Are Collecting Data to Understand Customers", https://medium.com/@melodyucros/10-sneaky-ways-companies-are-collecting-data-to-understand-customers-be0b9089d54a , Jan 12, 2018