IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Privacy preservation and Privacy by Design techniques in Big Data

M. Suresh Babu[1*], Mohammed Irfan[2], Suneetha V[3]

[1,2]Dept. of CSE, K.L. University – off Campus – Hyderabad, India
[3]Rayalaseema University, Kurnool, India

**Corresponding Author : principaliis@rediff.com, Tel : +91-9989988912*

DOI: https://doi.org/10.26438/ijcse/v7i4.588593 | Available online at: www.ijcseonline.org

Accepted: 13/Apr/2019, Published: 30/Apr/2019

Abstract- Big Data is a common term referring to a data revolution in information technology that makes it easy to collect, store and analyze user data online at relatively low costs. In simpler words, any human activity using technology leaves a *'digital exhaust'* or a trace data, a footprint. Broadly speaking, the big pool of all these collected footprints is called Big Data. However, it's not just a collection of these footprints but it also contains various other information like weather, train information, payments, etc. Generally these footprints may not have any apparent or obvious meaning, but they start to make sense when combined with other recorded datasets. This information could be processed using powerful analytic tools to give greater meaning and context to it while also enabling the system to 'predict' the unknown or missing information in the dataset. Today, we are already surrounded by a sea of ubiquitous sensors (sensors on your phones, punching access cards or swiping credit cards, etc). With each advancement, like the advent of the Internet of Things, coupled with the 'smartphone revolution' linking more and more information to your social media accounts, it is getting easier to gather more information and make sense of it. In this paper we discussed pseudonymization and privacy by design as the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information.

Keywords : Ubiquitous, Pseudonymization, privacy by design.

I. INTRODUCTION

Understanding the Basics of Data Protection

In 2018, the GDPR established a set of guidelines for managing the collection and storage of consumer and proprietary data. Much of it pertains to personal information provided by individuals to an entity. That entity may be a banking institution, insurance company, investing service, or even a health care facility. The primary goal is to ensure adequate protections are in place so that an ill-intentioned third party can't exploit the personal information of those organizations' employees, clients, and patients.

The key areas of data security:

- Explicit consent to collect and maintain personal data
- Notification in the event of a data breach
- Dedicated data security personnel within the organization
- Data encryption that protects personal information in the event of a breach
- Access to personal information for review of accuracy (integrity), and to set limitations on the intended use. While there has been pushback about some of the provisions within the Data Protection act (especially the need for additional data security personnel outside of the usual IT team), many organizations have been eager to adopt the measures. After all, being GDPR compliant can decrease the risk of a breach and would prove helpful if lawsuits resulted after a breach. In this paper we discuss about the Appropriate security, Implementation of Privacy by design in IoT, legal structure, Pseudonmyzation, Data minimization and Safe guard techniques in ubiquitous environment.

1.1 Appropriate Security

There is an ongoing discussion about what represents adequate and appropriate security in terms of data protection. To some degree, the exact approach to security will vary, based on the type of organization involved and the nature of the data that is collected and maintained. Even so, there is some overlap that would apply in every case. Compliance involves identifying and reinforcing every point in the network where some type of intrusion could possibly take place. Using Artificial Intelligence technology to reinforce points of vulnerability while also monitoring them for possible cyber attacks is another element. Even having an escalation plan in place to handle a major data breach within a short period of time is something any organization could enact. One point that is sometimes lost in the entire discussion about GDPR security is that the guidelines set minimum standards. Entities are free to go above and beyond in terms of protecting proprietary data like customer lists. Viewing compliance as the starting point and

continuing to refine network security will serve a company well in the long run.

II. IMPLEMENTATION OF PBD (PRIVACY BY DESIGN) IN IOT.

Both the terms 'Internet of Things' (IoT) and 'Privacy by Design' (PbD) were coined back in the 90s. The original idea behind PbD is to weave privacy into the very fabric of IT systems, networked infrastructure, business processes and design specifications; for that to happen successfully in the context of IoT, manufacturers of Internet-connected devices need to build privacy into their products from the ground up and at the outset of the developing process. In essence, the PbD is based on adherence to the 7 Foundational Principles of Privacy by Design:

7 Foundational Principles of PbD

Principle 1: Proactive not reactive: preventative not remedial
Principle 2: Privacy as the default setting
Principle 3: Privacy embedded into design
Principle 4: Full functionality: positive-sum, not zero-sum
Principle 5: End-to-end security: full lifecycle protection
Principle 6: Visibility and transparency: keep it open
Principle 7: Respect for user privacy: keep it user-centric

According to Dr. Ann Cavoukian – the founder of the Privacy by Design concept – explained in a 2016 report that "by embedding or coding privacy preferences into the technology itself, in order to prevent the privacy harms from arising," the PbD will achieve its goal to protect personal data and privacy at all stages of a product's development process. Nowadays, IoT is on the verge of becoming ubiquitous. San Jose, California has plans to create a smart city that will use transit vehicles and an infrastructure full of smart sensor appliances and technology with the ultimate goals of improving of safety, mobility and optimization of the transit system. The creators of this project claim it will deliver the "smart city" experience in a safest and most user-friendly way. Do they plan, however, to achieve that through the PbD approach?

Fig 1 : IoT devices in Ubiquitous environment

Cyberattacks against smart infrastructure do not remain in the sphere of science fiction; on the contrary – there have already been cases of compromised cameras, printers, weighing scales, doorbells, home routers and even connected fish tanks. Two examples of IoT products that have well-documented security issues – the lack of encryption and weak authentication mechanisms – are D-Link cameras and TP-link Smart Plugs.

Due to the boom of smart technology, the attack vectors continue to increase at a rapid rate. This is a clear illustration of the old maxim: when everything is connected, the network is only as strong as its weakest link. For IoT devices to be secure, one should protect hardware, software and connectivity. Unfortunately, most of the smart objects are not designed with strong (or any) security features built into their system. Limitations of IoT devices – e.g. insufficient processing power, memory and battery storage – constrain their capabilities to process information at a higher rate. Without serious consideration of the important matters of privacy and security of connected objects, there will be more botnets and more security breaches. Fortunately, the security firms seem to understand the gravity of the problems in question, as Gartner envisages worldwide spending on IoT security hardware, software, and services to reach $3.1 billion in 2021 (to make comparison: this figure is $1 billion in 2018).

The fact of the matter remains that in spite of knowing about privacy concerns regarding our personal data, the perception of 'what we gain' is more tangible and hence overpowers the realization of what are we trading off. For example, most people installed Truecaller at the pretext of knowing unknown callers (a tangible outcome), however, its ask for permission to access their contacts was probably overlooked as the outcome of that wasn't immediately comprehensible (hence not tangible). Personalization, when coupled with high privacy assurance, creates a significant positive association. It makes users more willing to share personal information and adopt a web based service. This highlights the importance of the user experience in terms of the perception of 'being in control' and the assurance for privacy. Which of the below messages are more willing to respond with an 'allow'?

2.1 Legal Structure of the Notion of PbD

With its Article 25 titled "Data protection by design and by default," the European Union's General Data Protection Regulation (the GDPR) adopted this notion officially, thus transforming it from a recommended best practice into a mandatory rule. Not only the EU was eager to mandate PbD – California Senate Bill 327 was introduced in the California Senate in April 2017, pursuant to which Web-connected devices should have built-security features appropriate to the nature of the device and the information it collects contains or transmits. Unfortunately, as mentioned before, privacy

and security are very poorly implemented in IoT product and service development, and these matters are handled often as an afterthought. In 2017, the Cyber Shield Act was introduced in the U.S. to remedy this problem.

Let's go back to the law that actually applies the said approach at the moment. Art. 25 of the GDPR makes mention of which methods data controllers/processors may choose to use in order to apply the PbD approach: "pseudonymization," "data minimization" and other "necessary safeguards [that] protect the rights of data subjects." Nevertheless, as the line "Such measures could consist, inter alia, of…" (Recital 78) suggests, this list is not, by all means, exhaustive.

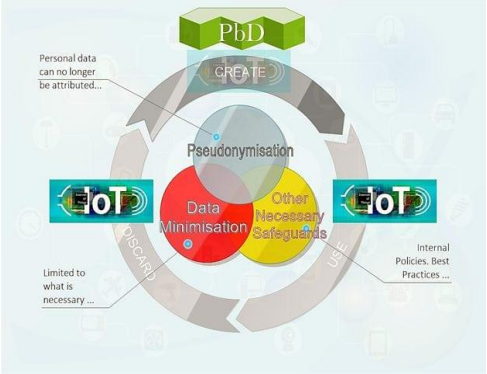

Fig 2 : Best Practices Related to Embedding PbD into IoT

2.2 Pseudonymization

It is important to be noted that anonymous data does not fall within the scope of the GDPR; hence, if you are able to completely remove all identifiers from personal data, it will not be deemed personal any more within the meaning of the EU data protection law.

Art. 4(5) of the GDPR

"'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;"

Although anonymity is often preferable, it is not always practical, because it runs counter to the principle of accountability. Pseudonymization, however, could bring about means that will strike a balance between anonymity and accountability. Pseudonymization utilizes a random identifier that secretly links to a person instead of a person's identity. A criminal would be incapable of directly identifying a data subject without linking the pseudonymization data to other sets of data stored and protected separately. In essence, this technique gives organizations the freedom to continue to process personal data under certain circumstances, as it protects individuals' right to privacy rather well.

2.3 The 7 De-Identification Techniques of WP29

1. Noise Addition: identifiers are expressed imprecisely (i.e., weight is expressed inaccurately +/- 10 kg).
2. Substitution/Permutation: identifiers are shuffled within a table or replaced with random values (i.e.a specific blood type is replaced with "Magenta").
3. Differential Privacy: identifiers of one data set are compared against an anonymized data set held by a third party with instructions of the noise function and acceptable amount of data leakage
4. Aggregation/K-Anonymity: identifiers are generalized into a range or group (i.e. age 43 is generalized to age group 40-55)
5. L-Diversity: identifiers are first generalized, then each attribute within an equivalence class is made to occur at least "L" times. (i.e. properties are assigned to personal identifiers, and each property is made to occur with a dataset, or partition, a minimum number of times).
6. Pseudonymization – Hash Functions: Identifiers of any size are replaced with artificial codes of a fixed size (i.e. blood type 0+ is replaced with "01", blood type A- with "02", blood type A+ is replaced with "03" etc).
7. Pseudonymization – Tokenization: identifiers are replaced with a non-sensitive identifier that traces back to the original data, but are not mathematically derived from the original data (i.e. a credit card number is exchanged in a token vault with a randomly generated token number).

Perhaps the biggest obstacle to effective pseudonymization is the difference in standards and proprietary technology of smart apparatuses.

2.4 Data Minimization

Art. 5(1) (c) of the GDPR

"Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');"

Data minimization is an essential element of the PbD concept, which requires services/applications in the realm of the IoT technology to process only the minimum amount of information necessary for the fulfillment of the particular service/function/transaction. The principle of data minimization can reduce both the size of the information IoT devices collect/process and the data retention period. Presumably, this would also reduce the chances of data handling issues (such as any kind of data misuse) or information theft.

Privacy enhancing technologies (PETs) that apply this principle can be designed not collect or store any personal information/personal identifiers (e.g. search history, search

terms, IP addresses). Ixquick (now StartPage), Unbubble, Disconnect and DuckDuckGo are excellent examples of such PETs. All tools that erase digital footprints – web browser cache, cookies, browsing history, address bar history, typed URLs, auto complete form history, saved passwords, search history, recent documents, temporary files, recycle bin and more – may be used to achieve the same effect in the context of Internet-enabled devices.

1. Collect only the fields of data necessary to the product or service being offered
2. Collect as little sensitive data as possible.

III. OTHER NECESSARY SAFEGUARDS AND BEST PRACTICES

The 2017 Verizon Data Breach Investigations Report, attested to the fact that 81% of hacking-related breaches happened due to weak passwords and 43% involved phishing – both of which are attacks that exploit the human factor. In addition, IT administrators are often failing to maintain best practices with respect to the IT infrastructure of which they are in charge. Perhaps the only feasible solution to correct such negligent behavior is to embrace privacy right from the outset. Manufacturers of connected products need to consider privacy at all times if their products process personal data. Usually, it is difficult and expensive to add privacy to a product at a later stage or, even worse, reengineer it following a failure. Widespread vulnerabilities like Heart bleed and Shellshock continue to plague IoT products. For that reason, it is essential to plan for future upgrades to device software. Unfortunately, many smart products are unpatchable.

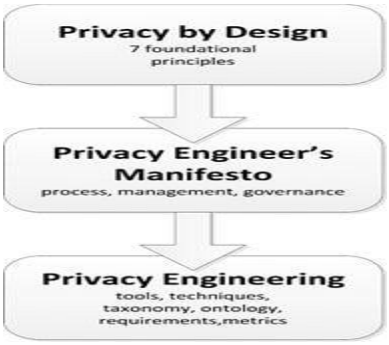

Fig 3 : Privacy by design flow

PbD embedded into connected objects also includes the presence of cyber-hygiene best practices, such as:

- **Security transmission protocols and encryption techniques for data in transit and at rest.** Protocols such as HTTPS and SSH are created to support encryption and strong authentication; unfortunately, the majority of IoT objects today can't use these features due to various inherent technical constraints.
- **Proper authentication** controls**, limiting permissions (assigned on a need-to-know basis).** Having usernames and passwords for every device is simply not feasible in an industrial environment. Alternative mechanisms, such as block chain, could solve the problem of trust and identity between smart objects. White listing of IoT clients may also prove useful in these situations. For critical communications, especially those that convey sensitive data, authentication and encryption measures are imperative for optimal protection, but providing a checksum or signature to allow the integrity of the data to be verified can be a recommended best practice with additional value to privacy and security.
- **Options to allow privacy/security default settings to be changeable.** This should include even hazardous services with a proven track record of creating vulnerable environments. For instance, many devices come from the factory equipped with non-essential services – Telnet or FTP, among others – that also pose high risk to users.
- **Training company staff** in privacy and data security best practices.
- **Application containerization**, where apps are installed in a contained environment (akin to virtual machines), could be beneficial privacy- and security-wise as well.

Some practical approaches that may facilitate the implementation of the PbD idea into a real and workable privacy shield are backend isolation, data separation, segregation, redaction and data transform techniques that remove personally identifiable information. It is advisable IoT products to have a button to switch off the "connectivity" function so that consumers can use them as regular products (e.g., from a connected plug into a regular one).

IoT products should undergo vigorous standard security testing, such as code analysis and ethical hacking, but also testing that specifically targets the effectiveness of the privacy-enhancing mechanisms. Data controllers need to vet data processors, vendors and other parties to know whether their cyber hygiene best practices live up to their expectations. Probably the most famous case of such a cyber attack was the one against Target, where the malicious actors gained control over the HVAC system of the company supplying Target.

When developers take into consideration the development of a product at the earliest stages, they should perform a thorough risk assessment and full analysis of potential attack vectors. The key to implementing PbD is the Data Protection Impact Assessment (DPIA). A DPIA is a process that

evaluates the risks associated with processing of particular personal data when it "is likely to result in a high risk to the rights and freedoms of natural persons." Although a DPIA is required only for companies categorized as high-risk, it is an integral part of the PbD approach. It should be carried out "prior to the processing in order taking into account the nature, scope, context and purposes of the processing and the sources of the risk."

There is a U.S. equivalent of the DPIA called cybersecurity disclosures, which are required by the Securities and Exchange Commission (SEC). In short, companies are obligated to discuss information security risks and incidents. Furthermore, the FTC is an advocate of the risk-based approach, which should take root from early stages through means of drafting a full inventory of the type and variety of personal information collected and subsequently understanding of data flows throughout the entire life cycle of all data sets. With respect to this point, the FTC noted:
"An evolving inventory serves triple duty: It offers a baseline as your staff and product line change over time. It can come in handy for regulatory compliance. And it can help you allocate your data security resources where they are needed most." To fully realize the potential of the PbD idea, enterprises should collect, map the flow, and analyze the data they handle.

PbD is, in fact, not only a responsibility of developers or similar people closely engaged with the manufacturing process. Everyone within the organization, from the software engineers to the marketing teams that make use of the applications, should be committed to the PbD.

You know the motto: "Security is everyone's responsibility." Privacy and security go hand in hand, so "security by design" could complement the value of the PbD. As Isabelle Noblanc wrote at PYMNTS.com: "Security isn't the hot sauce you add on the side. It's a key ingredient to any system, and it's something IoT manufacturers need to think about from the very beginning."

Transparency + Fairness + User Control = Trust

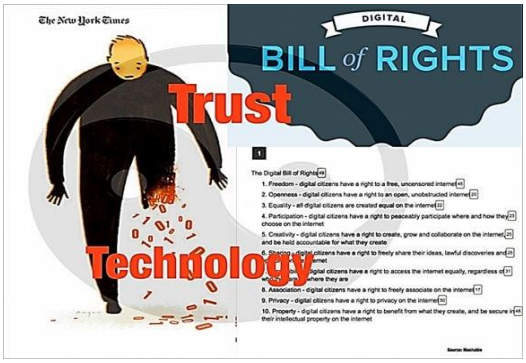

Fig 4 : Trust and Technology

The notion of PbD alone may not be enough to promote privacy without a working regimen on how service providers are to obtain consumers' meaningful consent. Terms of service should be designed to prevent service providers from using the personal data of their customers, unless opt-in consent has been obtained in advance. It is important for IoT companies, as far as privacy is concerned, to vest their users with the power to control their personal data, and that is usually done based on the principle of "transparency."

Fig 5 : Privacy concerns in Big Data

An ESET research team investigated privacy concerns associated with some popular IoT products. According to their findings, voice-activated intelligent assistants seemed to raise most concerns, since there is a greater probability for interception of digital traffic by cybercriminals, over sharing of data among service providers may not be uncommon, and the overall state of data protection is not up to par. Smart technology brings about convenience, but the price the users pay is in their personal data, which is mined, analyzed and sold. Unfortunately, many companies build their businesses around data mining and analysis, and are therefore rather reluctant to adopt PbD practices. But at least two benefits are beginning to arise through the implementation of PbD: the user is assured strong privacy and control over their own information, and organizations gain competitive advantage.
Trust is quickly becoming an important asset in the digital ecosystem. It has become a form of currency, as wary customers are now on the lookout for companies who have demonstrated a commitment to maintaining security and privacy. Perhaps PbD can be the cornerstone on which IoT companies build their trust relationships with their clients.

IV. CONCLUSION

Big Data has the potential to generate enormous value to society. In order to ensure that it does, opportunities to enhance privacy and civil liberties are best conceived early on. In this paper we have explored the emergence of Privacy by Design systems as an emerging capability with an unprecedented ability to integrate previously diversified data and in some cases, data about people and their daily lives.

The use of advanced analytics has made it possible to analyze large data sets for emerging patterns. It is increasingly apparent, however, that these techniques alone will be insufficient to manage the world of Big Data especially given the need for organizations to be able to respond to risks and opportunities in real time. Next-generation capabilities like sense making offer a unique approach to gaining relevant insights from Big Data through context accumulation. While these new developments are highly welcome, building in privacy-enhancing elements, by design, can minimize the privacy harm, or even prevent the privacy harm from arising in the first place. This will in turn engender greater trust and confidence in the industries that make use of these new capabilities. The dynamic pace of technological innovation requires us to protect privacy in a proactive manner in order to better safeguard privacy within our societies. In order to achieve this goal, system designers should be encouraged to practice responsible innovation in the field of advanced analytics. With this in mind, we strongly encourage those designing and building next generation analytics of any kind to carry out this work while being informed by Privacy by Design as it relates to personally identifiable data.

REFERENCES

[1]. Lee Chung, H.; Cranage David, A. 2010. *Personalisation-privacy paradox: The effects of personalisation and privacy assurance on customer responses to travel websites.* Elsevier. http://www.elsevier.com/locate/tourman

[2]. Yanying Gu, Anthony Lo, 2009. *A Survey of Indoor Positioning Systems for Wireless Personal Networks*. IEEE Communications Surveys & Tutorials, Vol. 11, №1, First Quarter.

[3]. Manyika, J., et. al. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Online: http://www.mckinsey.com/Insights/MGI/Research/Technology_andInnovation/Big_data_The_next_frontier_for_innovation.

[4]. Tene, O., and Polonetsky J. (2012). Privacy in the age of big data: A time for big decisions. Stanford Law Review 64, 63.

[5]. Commission Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), COM (2012) 11 final (Jan. 25, 2012). Online: http://ec.europa.eu/justice/ newsroom/data-protection/news/120125_en.htm.

[6]. Gantz. J., and Reinsel. D. (2011). Extracting value from chaos. IDC. Online: http://www.emc.com/ collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[7]. Jeff Jonas and Lisa Sokol (2009), "Data finds data," in Segaran, T., and Hammerbacher, J. (eds.), Beautiful Data The Stories Behind Elegant Data Solutions, O'Reilly Media. p. 105.

[8]. Jonas, J. (Oct 11, 2010). On how data makes corporations dumb. GigaOm. Online: http://gigaom. com/2010/10/11/jeff-jonas-big-data/.

[9]. Marsella, A., and Banks, M. (2005). Making customer analytics work for you! Journal of Targeting, Measurement and Analysis for Marketing. 13(4), 299-303.

[10]. Jonas, J., and Harper, J. (2006). Effective counterterrorism and the limited role of predictive data mining. Policy Analysis. CATO Institute, Washington, DC, 584, 1-11. 13 Jonas, J. (2009). Data finds data. Online:http://jeffjonas.typepad.com/jeff_jonas/2009/07/data-findsdata.html

[11]. Privacy and Security by design is a crucial step for privacy protection., Least Authority Kingsmill, S. & Cavoukian, A. *Privacy by Design: Setting a new standard for privacy certification*

[12]. Maple, C., *Security and privacy in the internet of things*, Taylor and Francis Online

AUTHORS PROFILE

Dr. M. Suresh Babu, received Bachelors Degree from Sri Krishnadevaraya University, Master of Computer Applications from Osmania University, M.Phil from Bharathiar University and PhD in Computer Science from Sri Krishnadevaraya University. He is having 22 years of academic and administrative experience. He worked as Professor, Principal and Chairman Board of Studies, Member BoS for various autonomous colleges. He has contributed more than 96 papers in various national and International Journals, Conferences, and Symposiums. At present, he is working as Professor in Department of CSE, K.L.University - Hyderabad off Campus.

Mr. Mohammed Irfan, received Bachelors Degree in Computer Applications (BCA) from Ghasidas Vishwavidyalaya University, a Central University in India in 2001 and Master Degree in Computer Applications (MCA) from Osmania University, accredited by NAAC, A+ Grade University in India in 2005. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and Engineering, in KL University Hyderabad since 2018, also worked as a Lecturer in Abha, SaudiArabia past 12 years. He has publications in refereed international journals and conferences including IEEE and it's also available online. His main research work focuses on Data Mining, IoT , Cloud Security and Privacy and Big Data Analytics. He has 13 years of teaching experience and 5 years of Research Experience.

