DATA ANONYMIZATION PROCESS CHALLENGES AND CONTEXT

Hassane Tahir[1] and Patrick Brezillon[2]

[1]Phd Computer Science, Member of ACM, France
[2]LIP6 - Sorbonne Université

ABSTRACT

Data anonymization is one of the solutions allowing companies to comply with the GDPR directive in terms of data protection. In this context, developers must follow several steps in the process of data anonymization in development and testing environments. Indeed, real personal and sensitive data must not leave the production environment which is very secure. Often, anonymization experts are faced with difficulties including the lack of data flows and mapping between data sources, the non-cooperation of the database project teams (refusal to change) or even the lack of skills of these teams present due to the age of the systems developed by experienced teams who unfortunately left the project. Other problems are lack of data models. The aim of this paper is to discuss an anonymization process of databases of banking applications and present our context-based recommendations to overcome the different issues met and the solutions to improve methodologies of data anonymization process.

KEYWORDS

Context, Data Anonymization, GDPR, Personal Data, Process of Anonymization, Sensitive Data.

1. INTRODUCTION

Anonymization is one of the technical measures that can be implemented by data controllers. According to the CNIL [7], this is an operation which consists of using a set of techniques, so that to make impossible, in practice, any identification of the person by any process whatsoever and in an irreversible manner. Once anonymized, the data is no longer subject to the GDPR (General Data Protection Regulation) since it is no longer personal. Anonymization allows data to be retained for defined processing, without it being possible to attribute it to the person concerned. It is useful for retaining data beyond its initial retention period, but also for reusing it, publishing it, etc. However, one must be vigilant on this point and ensure that the anonymization process is truly effective. If anonymization is carried out well, it makes it possible to exploit the data, which are at the origin of personal data, while limiting the risks of violating the privacy of individuals.

For companies processing "personal and sensitive data", this must not be accessible in development and test databases (Dev/Test), to comply with GDPR regulations and protect company against possible violations [11]. Indeed, by nature, a basic Dev/Test environment must be open, and present few constraints so as not to harm the work of designers, developers, and testers. The data on these environments must be as close as possible to the Prod data to preserve their consistency. The best solution to protect data while meeting these requirements is to anonymize sensitive data in non-production environments.

DOI: 10.5121/ijdms.2023.15601

Today there exists many anonymization tools commercial and open sources. They can automate the process of identity protection and are generally based on methods such as Pseudonymization and Generalization. Data anonymization tools can be applied to various use cases, including Software testing, Marketing analytics, medical research, Business performance. Our work is based on Informatica TDM (Test Data Management) masking tool. In this paper, we present a case study on the data anonymization process using Informatica TDM to mask personal and sensitive data of bank customers with car loans.

2. PERSONAL DATA PROTECTION

With the exponential growth of the data economy, businesses and organizations are collecting more personal data than ever, from large different sources, including e-commerce, government and healthcare sources, and social media. Therefore, the risk of personal information being accessed and misused is greater than ever due to the ever-increasing amount of data collected and stored. When personal information is violated, a breach of organizational security and a breach of trust for customers will occur and lead to attacks and broad privacy violations, including breach of contract, discrimination, and theft identity.

Data analysis in the anonymization process includes a very important activity named Personal data discovery. CNIL (French regulatory institution for personal data) defined personal data as: “Any information relating to an identified or identifiable individual; an identifiable person is one who can be identified, directly (i.e. last name, first name) or indirectly (i.e. phone number, license plate of a vehicle, an identifier such as social security number, postal code or email address, but also voice or image)” (CNIL 2021). A person can be identified by a single piece of data such as “Name” or by data correlation (i.e. a girl living at a known address in the city, born on given day, and being a member of a sport association). On the other hand, company contact details (i.e. Company ABC with its postal address, telephone number, contact email “company_abc@email.fr”) are not personal data. Sensitive data is a special category of personal data that is related to racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership, as well as genetic data, biometric data for the purpose of uniquely identifying a person or an individual, data regarding health, sexual life or orientation of a person.

Data anonymization has a very important role in protecting privacy by preventing the exposure, and exploitation, of people’s personal and sensitive information. It alters personal or sensitive data so that it can’t easily be linked to a specific individual or business organization. This will minimize the risk of *re-identification* to comply with data privacy regulation and heighten security. Hence, data anonymization addresses the privacy concerns associated with data sharing by making it difficult to re-identify individual information from the datasets. The anonymization process includes data deletion, or data masking, of any Personally Identifiable Information (PII). Examples of personal data are names, addresses, telephone numbers, passport details, or Social Security Numbers. To protect data, values are replaced or removed, by using cryptographic techniques, or adding random noise.

3. RELATED WORK IN DATA ANONYMIZATION

3.1. Anonymization Techniques and Algorithms

There are many data anonymization methods including generalization, suppression, perturbation, anatomization, permutation. The generalization transforms the original values into less specific but semantically consistent values during anonymization process. This can be achieved by replacing the original values with others semantically similar but less specific [22]. Suppression

is an operation that hides an original value with a special value (i.e., '*' or '?'.). It applies to quasi-identifier attributes which can be either categorical or numerical [26] . The perturbation method allows replacing the original data with synthetical data. In other words, the values of the quasi-identifier attributes are replaced by fictitious values. The first advantage of this technique is the preservation of the statistical results. The second important advantage is that an individual cannot be identified by an attacker by linking quasi-identifier attribute values or by accessing the individual's sensitive information. Indeed, the values are synthetic, and it is not possible to publish the original information about an individual [10]. In the anatomization technique, the attributes are separated into two tables: one table for the quasi-identifier attributes and another for sensitive attributes [1]. The main advantage of anatomization is that the data in the sensitive and quasi-identifier tables is not modified. An improved version of the anatomization technique called "permutation" involves applying a random permutation of the values before publishing the data. Permutation can ensure strong privacy preservation and good data utility [17].

The most important anonymization algorithms are k-anonymity, l-diversity, t-closeness, δ-disclosure privacy, and δ-presence. The generalization and/or suppression techniques are used with the k-anonymity algorithm. It works on the principle that if you combine data with similar attributes, you can obscure identifying information about any individual contributing to that data. Generalization and/or suppression techniques can be performed on quasi-identifier attributes by applying the k-anonymity algorithm [23]. The value k defines the level of confidentiality and is linked to the loss of information. Thus, the higher the value of k, the higher the privacy value and the utility of the data is even lower [6]. Like k-anonymity, l-diversity can't guarantee absolute privacy protection for anonymized data and it's much more difficult to implement than k-anonymity since identification. In addition, protection of sensitive attributes can only work if there are at least L distinct values for each sensitive attribute in the dataset. The major difference between the k-anonymity algorithm and the l-diversity algorithm is that the former works on quasi-identifier attributes while the latter works on sensitive attributes [26] . According to Li, Li, and Venkatasubramanian [18], the l-diversity algorithm had some vulnerabilities, and that l-diversity was not even necessary to prevent attribute disclosure. They therefore propose a new notion of privacy and consequently a t-closeness algorithm which leads to less information loss than l-diversity and k-anonymity, but all techniques still lead to a significant loss value. In addition, as the number of attributes increases in the dataset, the information loss also increases [3]. On the other hand, Brickell & Shmatikov [5] noticed that k-anonymity algorithms do not guarantee high privacy to users. Hence, they introduced δ-disclosure privacy algorithm based on privacy metrics considering syntactic properties of the anonymised databases. Nergiz, Atzori and Clifton [20] showed in their work that their algorithm is a good approach when existing algorithms are not the most appropriate.

3.2. Data Anonymization Tools

An anonymization tool is software that allows users to anonymize a large amount of data quickly and easily and can prevent them from introducing errors into the anonymized data set. It is based on the anonymization techniques and algorithms.

The open-source anonymization tools could be used by everyone. They are based on different algorithms, techniques and might have performance discrepancies in the same environment. Sartor [24], in collaboration with Aircloak GmbH, a company who's the main concern is the responsible use of personal data, named ARX Data Anonymization Tool, Amnesia, sdcMicro and μ-ARGUS as the top data best anonymization tools in 2019. These are therefore not the solution for big companies, because data anonymization is a very complex process and must be carried out by use case and not dataset by dataset. ARX is used to anonymize sensitive personal data [4]. It is used in various contexts such as research projects, commercial big data analytics

platforms, clinical trial data sharing, and training. It supports most of the anonymization algorithms. Amnesia anonymization tool is a software used locally to anonymize personal and sensitive data [2]. It currently supports k-anonymity and km-anonymity guarantees. DOT-Anonymizer is a tool that preserves the confidentiality of test data by hiding personal information [9]. This tool anonymizes personal data while preserving its format and type. µ-ARGUS is an anonymization tool based on the R programming language. ARGUS means Anti Re-identification General Utility System. It can create secure microdata files. The tool uses different statistical anonymization techniques such as top and bottom coding, micro aggregation, adding noise, randomization, local suppression, and global recoding. µ-ARGUS can also be used to create synthetic data [12]. sdcMicro is another open-source data masking tool published in May 2018. It is an R-package used to generate anonymized microdata such as public and scientific use files.

Other anonymization commercial tools include IBM InfoSphere Optim Data Privacy [14], Informatica TDM [15], K2view [16], Dataprof [8]. Persistent Data Masking helps secure sensitive data through anonymization and encryption in software development, support of testing, analytics, and non-production environments. It enables scalability, management, and connectivity across traditional databases, Apache Hadoop, and cloud environments, while adhering to consistent data anonymization policies across organizations and companies with a single audit trail. One of the Informatica TDM Architecture will be described in the case study.

All the above anonymization tools are very useful for protecting personal data. Unfortunately, the process of anonymization remains a hard and complex process requiring more and more working loads and often exceed the initial estimated time to successfully complete the anonymization. We will point some of challenges that should be considered to improve the anonymization tools.

3.3. Sensitive and Personal Data Discovery Tools

Data discovery tools allow someone to examine all database sources to find sensitive and personal data and then classify it. This will make it possible to generate precise results, with few false positives, but it is problematic to do things well. However, someone tasked with reviewing many database tables may be careless and miss some data that should be reported. Usually, people are bad at doing repetitive and mundane tasks.

Data discovery procedures have been developed to enable the automation of repetitive actions aimed at detecting personal and sensitive data based on different criteria and techniques such as the use of regular expressions and dictionary-based rules (i.e. cities, first names, etc.). Once the discovery procedure is established, it can be applied in many other uses in addition to respecting data privacy and security. This can help locate relevant information in archives by applying a different set of rules.

The following are some related research works in data discovery. Securiti [25] is an AI-based tool that allows organizations to discover sensitive data stored across multiple platforms (i.e., Cloud, SaaS, and on-premises environments). This also helps protect this data and automate many privacy functions. We can also mention another tool like IBM Security Guardium [13] which allows someone to analyze, discover, classify, and control access to sensitive data. Finally, SQL Server Management Studio (SSMS) can be cited as a tool for discovering, classifying, and labeling columns containing sensitive data in a database, as well as visualizing the current classification status of the database and to export reports [19]. Many other related research work discussions can be found in literature, but we cannot list all of them.

4. CASE STUDY

4.1. Context

The context of the project is the anonymization of all the personal and sensitive data in the databases of all applications of a bank and its national and international subsidiaries. The work is carried out in collaboration with an anonymization team made up of developers within a bank's information security department.

4.2. Informatica Test Data Management (TDM) Masking Tool

Test Data Management (TDM) is an Informatica masking tool to manage nonproduction data in an organization. It can create a smaller copy of production data and anonymize sensitive data. Sensitive columns can be discovered in the test data to ensure that they are masked in the Dev/Test environment. TDM is also used to generate test data that does not contain sensitive production data. Multiple copies of application data can be created to be used for testing and development. Strict controls are maintained on production systems, but data security in non-production systems is not as secure. Therefore, having knowledge of sensitive columns in production databases, one must ensure that sensitive data does not appear in development environments. The TDM architecture implemented in our Anonymization environment is shown is Figure 1 where the masking environment can be accessed using a browser (i.e. IE or Chrome). The portal allows to create anonymization projects and rules. It also allows generating anonymization plans.

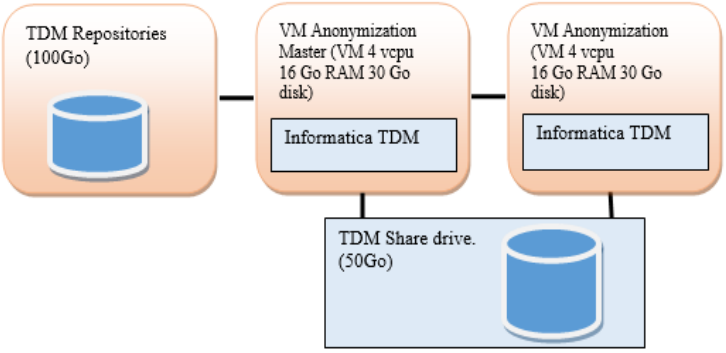

Figure 1. TDM Architecture.

4.3. Anonymization Phases

The process of Anonymization is composed of six phases as shown in the Figure 2. All requirements regarding the databases to be anonymized must be specified in phase 1. Then two clones of the source production database will be created in phase 2: One source clone database and one clone masked target database. In the third phase, sensitive data should be identified, and masking rules must be created in the Informatica TDM masking platform. Once all the anonymization plans are developed and executed (phase 4), by reading source data from the source clone database and then inserting the anonymized data into the target clone database. In phase 5, technical and functional tests will be performed to validate the masked data to import the anonymized data into the "Acceptance database" in the non-production environment. Finally, to secure the source data of the production database, the data must be deleted from the source and target clone databases.

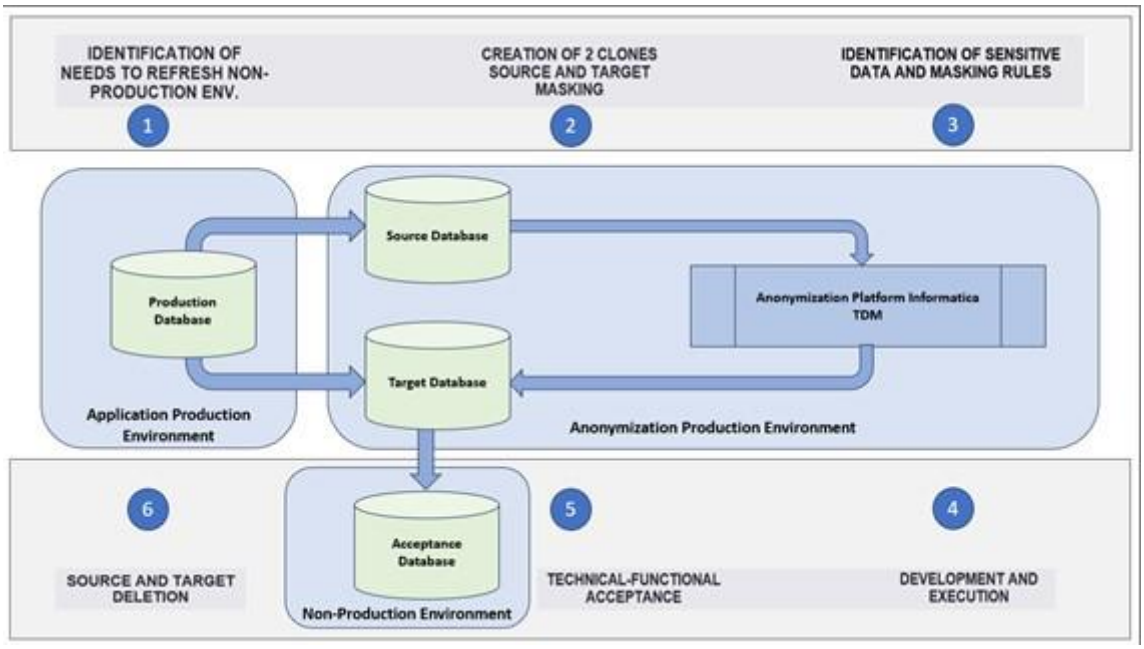

Figure 2. An example of the Anonymization Process.

5. CHALLENGES FACED BY DATA ANONYMIZATION EXPERTS

The process of anonymization can be very complex and time-consuming, particularly when dealing with huge amounts of data. Although automatic anonymization tools solve problems induced by manual data processing, there are certain challenges that need to be solved to avoid a risk of human error in the anonymization process, which could result in personal data being inadvertently disclosed. In this paper, we will not discuss the issues related to anonymization techniques but to the contextual elements acting on the anonymization process. Related work in contextualizing the well-established procedures and processes can be found in our previous work as in [27]. Three types of challenges can be distinguished when in the process of data anonymization, which are related to organizational context, functional or business context, and technical context. Each of these three contexts includes a list of related contextual elements [28], and the instantiation of these contextual elements can alert if the anonymization process can be short-circuited. Here, the contextual elements are clearly identified (i.e. the anonymization process is supposed to have a robust model), but the instantiation of one of them can play the role of a weak signal that is generally not detected.

5.1. Challenges Related to Organizational Context

Often in the analysis phase of the anonymization process, developers often need to interact with different actors to gather all information required to the technical implementation. Considering the user context can help improve the data anonymization. In our anonymization projects we have met many difficulties when working with some actors who cannot answer to our questions in the required forms. Sometimes actors don't master the data model of an application as well as the data flow between the different applications and they don't know whether their databases are used in a non-secure non-production (i.e. test) environment and if these databases contain personal or sensitive data. There are automatic discovery tools designed to support business analysts in specifying personal and sensitive data (contextual elements), but they are not always useful because they don't consider organizational context where contextual elements are

instantiated. In the next section we will present a context-based discovery tool used to improve the analysis phase of the anonymization process.

Application and business managers can efficiently help developers if they have a good level of business or functional knowledge of the application as well as mastering the data model. Other contextual elements include business area and the perenniality of the application. Figure 3 shows an example of different context regarding the business user knowledge about the model of data flows between the different databases within the development environment.

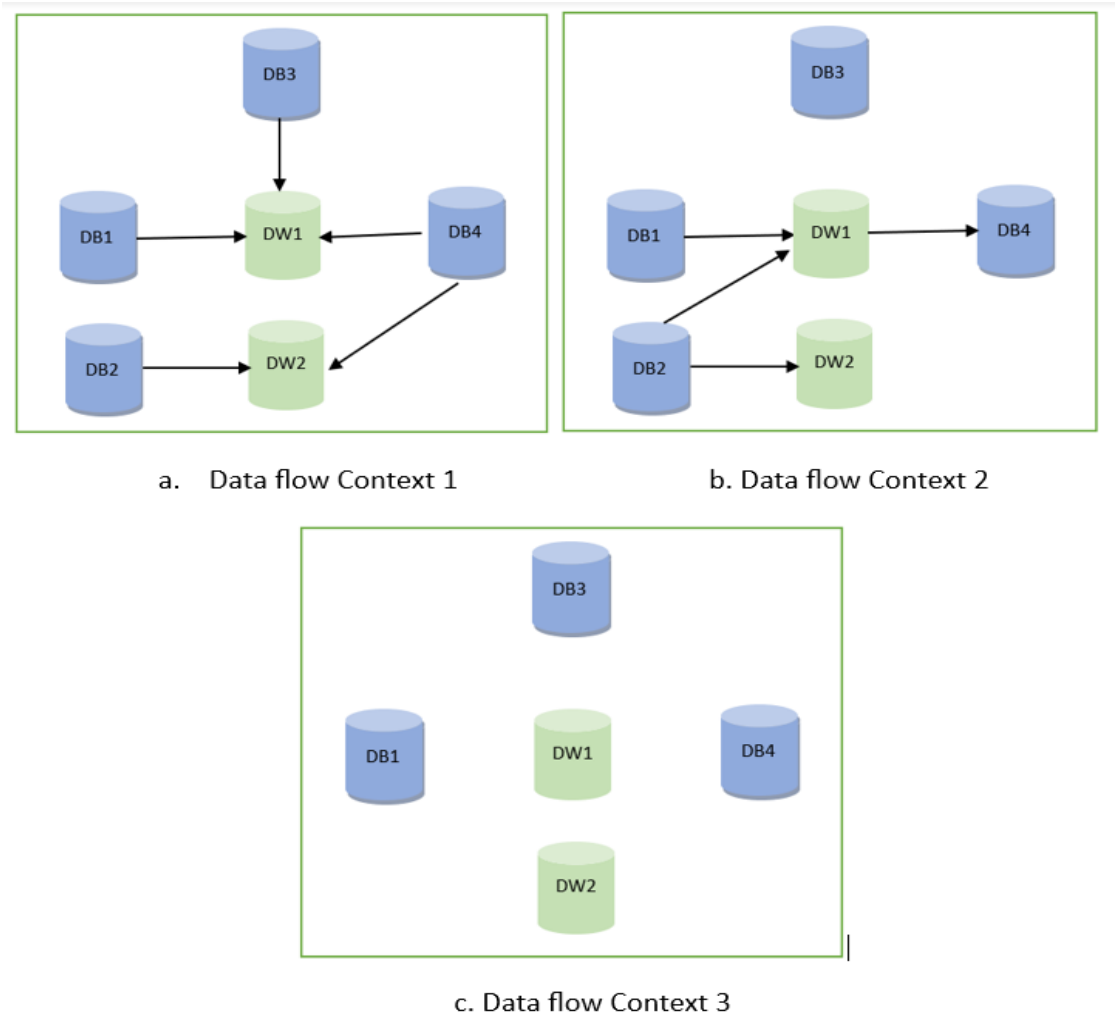

Figure 3. Different contexts about the business user knowledge

In the Data flow context 1, all the personal data stored in databases DB1, DB2, DB3 and DB4 must be anonymized. On the other hand, there is no need to anonymize data loaded into the two warehouses DW1 and DW2 which are already anonymized in the sources. In the Data flow context 2, the anonymization is the same as for context 1 except that DB4 which should not be anonymized as its data is loaded from DW1 which is anonymized. Finally in context 3, the business analyst doesn't know what the data flow between all the databases is. Therefore, what is the decision to be made? Do developers should anonymize all the databases? In this case, there will multiple anonymizations which may lead to data inconsistency.

5.2. Challenges Related to Functional/Business Context

The functional or business context include whether the application contains personal or sensitive data and the type of this data (i.e. health or sexual life, racial or ethnic origin, financial, etc…) and if the application contains HR, customer, supplier data.

5.3. Challenges Related to Technical Context

The technical context concerns knowledge of the structure of data, the database environment storing the data (i.e., production, development, testing, etc.), data flows, tools used to secure and protect the data, etc.

6. CONTEXT-BASED TOOLS FOR IMPROVING ANONYMIZATION

As discussed in the previous section, in addition to automatic anonymization tools, our approach is to consider context to improve the anonymization process. In our project, we have introduced a framework for a context personal data discovery. Contextual graphs software [21] can be used as a decision-making tool to assist business analysts in personal data classification when working on databases with too many tables. A contextual graph is a knowledge acquisition tool (KAT) that has been improved to generate the user's practices data, which will be useful in creating indicators that can be used to assist the user in reading practices in the graph. It is used by a business analyst to adapt the result classification of personal data (by the discovery tool) to the context about the meaning of the data stored in each field of the database table. The main components of the KAT contextual graph are CxG Editor, CxG Reader and an experience database. The main advantage of the proposed tool is to provide a uniform representation of knowledge, reasoning in context, progressive knowledge acquisition and practical learning. The architecture of the context-based personal data discovery tool is shown in the Figure 4. All details about our approach and the implementation can be found in our book chapter in [28].

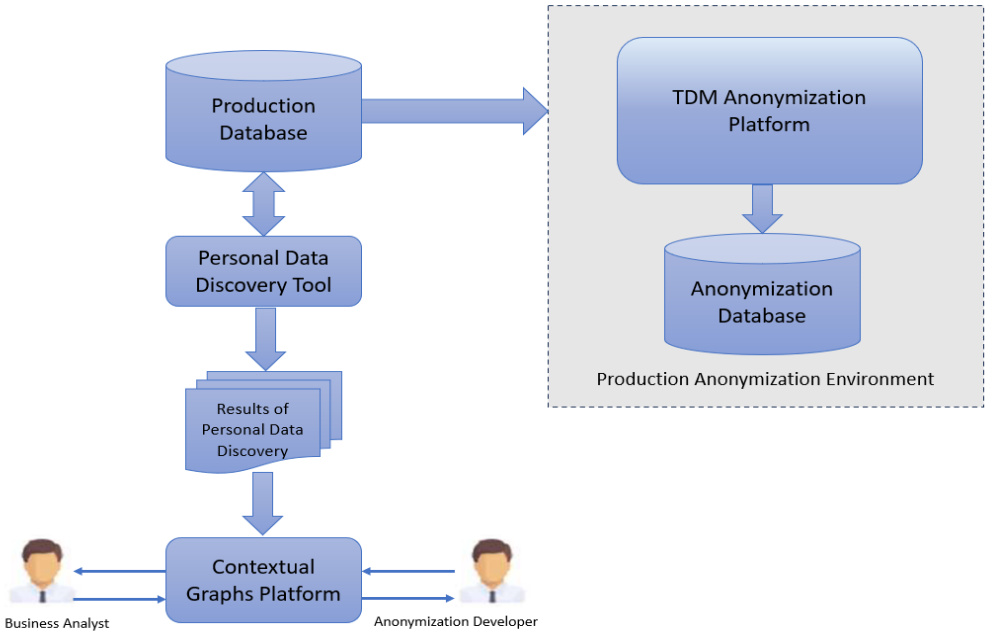

Figure 4. Architecture of a Context-based Personal Data Discovery Tool.

7. CONCLUSIONS

This paper presented some issues encountered in the data anonymization process of one of the important projects in the banking sector, as presented in a case study. We discussed some of the important anonymization algorithms and techniques as well as tools for discovering and classifying personal and sensitive data. We presented the challenges faced by anonymization experts which are linked to three types of contexts: organizational, functional/business, and technical. Finally, we presented an architecture for a context-based personal data discovery tool as a solution to some of the challenges discussed. This architecture is based on a Contextual Graphs approach which has already been applied to other fields, particularly in medicine and industry.

REFERENCES

[1] Almokbily, R. S., & Rauf, A. (2018). Anatomization through generalization (AG): A hybrid privacy-preserving approach to prevent membership, identity and semantic similarity disclosure attacks. In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-7).

[2] Amnesia Anonymization Tool - Data anonymization made easy (openaire.eu), https://amnesia.openaire.eu/, last accessed 2023/10/20.

[3] Arora, D. K., Bansal, D., & Sofat, S. (2014). Comparative Analysis of Anonymization Techniques. In International Journal of Electronic and Electrical Engineering (pp. 773-778). International Research Publication House.

[4] ARX - Data Anonymization Tool | A comprehensive software for privacy-preserving microdata publishing (deidentifier.org), https://arx.deidentifier.org/, last accessed 2023/10/20.

[5] Brickell, J., & Shmatikov, V. (2008). The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 70-78). Nevada, USA: Association for Computing Machinery.

[6] Brito, F. T., & Machado, J. (2017). Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações. In Jornadas de Atualização em Informática - Chapter: 3 (p. 40). Brasil: Sociedade Brasileira de Computação - SBC.

[7] CNIL, https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles, L'anonymisation de données personnelles, last accessed 2023/10/20.

[8] DATPROF, Test Data Management - DATPROF, https://www.datprof.com/, last accessed 2023/10/20.

[9] DOT Anonymizer, Data masking Tool (dot-anonymizer.com), https://www.dot-anonymizer.com/data-masking/dot-anonymizer/, last accessed 2023/10/20.

[10] Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Comput. Surv. Volume 42.

[11] General Data Protection Regulation GDPR, https://gdpr-info.eu/, last accessed 2023/10/20.

[12] Gramener.com, 10 Best Data Anonymization Tools & Techniques, https://blog.gramener.com/10-best-data-anonymization-tools-and-techniques-to-protect-sensitive-information/, last accessed 2023/10/20.

[13] IBM, IBM Security Discover and Classify, https://www.ibm.com/fr-fr/products/ibm-security-discover-and-classify, last accessed 2023/10/20.

[14] IBM InfoSphere Optim Data Privacy, https://www.ibm.com/products/infosphere-optim-data-privacy, last accessed 2023/10/20.

[15] Informatica, Informatica Test Data Management, https://www.informatica.com/content/dam/informatica-com/en/collateral/data-sheet/informatica-test-data-management_data-sheet_3234en.pdf, 2023/10/20.

[16] K2view, Data Masking Tools | K2View, https://www.k2view.com/solutions/data-masking-tools/, last accessed 2023/10/20.

[17] Li, D., Xianmang, H., Cao, L., & Chen, H. (2015). Permutation anonymization. Journal of Intelligent Information Systems.

[18] Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering (ICDE), 106-115.

[19] Microsoft.com, SQL Data Discovery and Classification,https://learn.microsoft.com/en-us/sql/relational-databases/security/sql-data-discovery-and-classification?view=sql-server-ver16&tabs=t-sql, last accessed 2023/10/20.

[20] Nergiz, M. E., Atzori, M., & Clifton, C. (2007). Hiding the presence of individuals from shared databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, 665-676.

[21] Pomerol J.-Ch., Brézillon P. and Pasquier L. (2002) Operational knowledge representation for practical decision making. Journal of Management Information Systems, 18(4): 101-116.

[22] Rao, P. S., & Satyanarayana, S. (2018). Privacy preserving data publishing based on sensitivity in context of Big Data using Hive. Journal of Big Data, Article number: 20.

[23] Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. technical report, SRI International.

[24] Sartor, N. (2019, February 4). Top 5 Free Data Anonymization Tools. Retrieved from aricloak: https://aircloak.com/top-5-free-data-anonymization-tools/

[25] Securiti.ai, Automate Data Access Governance for Sensitive Data, https://securiti.ai/products/access-intelligence/, last accessed 2023/10/20.

[26] Sharma, S., Choudhary, N., & Jain, K. (2019). A Study on Models and Techniques of Anonymization in Data Publishing. International journal of scientific research in science, engineering and technology, 84-90.

[27] Tahir, H. and Brézillon, P. (2013) Shared context for improving collaboration in database administration. International Journal of Database Management Systems (IJDMS) 5(2): 13-28.

[28] Tahir, H. and Brézillon, P. (2022) Context-Based Personal Data Discovery for Anonymization, in Modeling and Use of Context in Action, by Patrick Brézillon and Roy M. Turner, First published 2022 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc., ISTE Ltd 2022.