POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Investigation and Implementation of Dynamic Data Masking

Supervisor
prof. Alessandro Savino
Co-Supervisors
Vincenzo Cannata
Nicolò Maunero

Candidate
Alain Divin Bahizi

Academic Year 2023-2024

ACKNOWLEDGEMENTS

As I reflect on the journey that culminates in this thesis, I am profoundly aware that my achievements stand on the foundational support of my family and loved ones.
A special acknowledgment is reserved for my thesis supervisor, Nicolò Maunero, whose mentorship has been nothing short of inspirational. You have been an ideal mentor and supervisor. Your dedication to my development is deeply appreciated.
I extend my heartfelt gratitude to my friends and family, who provided unwavering support during the most challenging moments of this journey. Your belief in my potential helped me navigate through tough times.
A very special thanks goes to my mom and sister, whose love and belief in me have been the pillars of my strength. Without you, this achievement would still be a dream. Your sacrifices and faith in me have transformed this dream into reality.

To all of you, I dedicate this work.

CONTENTS

LIST OF FIGURES

ABSTRACT

This thesis investigates the adoption and implementation of data masking solutions within the financial services sector, particularly in response to the European Union's landmark 2016 regulation on data protection. Article 25 of this regulation mandates that organizations employ adequate technical and organizational measures to integrate data protection principles by design and by default. Data masking emerges as a critical strategy for achieving compliance, transforming the way businesses handle user data and setting a benchmark for data privacy practices worldwide.

The study is anchored in a practical experience gained through an internship at PwC Italy, aiming to implement a data masking solution that meets the specific GDPR compliance needs of a client.

The investigation covers the spectrum of data masking, distinguishing between static and dynamic data masking techniques, their implementation approaches, and how they cater to varying data protection requirements. Through the lens of significant data breaches, the thesis evaluates their impacts and legal consequences, underpinning the critical need for proactive data protection strategies like data masking.

A thorough analysis of regulatory compliance is presented, focusing on the legal mandates that organizations must fulfill concerning data protection. This discussion underscores the role of data masking within the regulatory framework, showcasing its significance in meeting data protection standards.

Given the array of available commercial tools for data masking, the thesis navigates the complexities of selecting appropriate solutions. It considers factors such as functionality, cost, and organizational integration, providing a nuanced view of the decision-making process involved.

The culmination of the thesis is a detailed examination of the data masking solution implemented during the author's internship. This section reveals the achieved outcomes, the technical intricacies of the solution, and the challenges encountered throughout the project's development.

By focusing on the application of data masking techniques not only for data anonymity but also for legal compliance with GDPR, this thesis contributes valuable insights into the evolving landscape of data privacy and protection. It underscores the pivotal role of data masking in contemporary data protection strategies, offering a comprehensive overview for organizations aiming to navigate the complexities of regulatory compliance and data security.

CHAPTER 1

INTRODUCTION

As our world becomes more and more connected, data breaches happen more frequently, not only in small business with no security specialists and budget but also for government and big businesses. In this chapter we are going to analyse and focus on data breaches, what causes them and what are their consequences in our society.

Picking from the real world scenarios like the first American Corporation data breach, the city of Calgary data breach or even the solar wind data breach, statistics shows that they happen more often than we might think so and could have devastating effects on your daily life.

Even though, it seems like data breaches are unavoidable or difficult to contain, there are solutions that can decrease considerably the risk of them. One of those solutions is Data masking. Data masking is a way of producing fake but convincing versions of the organization's data to protect it from unauthorized access.

From manipulation to shuffling, data masking uses different techniques to obfuscate the data within a database. This technique is very useful when you want your data to remain useful to other parties. We could take an example of developers who need to test their application to the database, they don't need the real sensitive data but the ones they can work with.

We will also discuss the legal aspects of data breaches, data masking is of greatest importance as masking becomes more and more important. Legislation surrounding data protection are in constant change. Later in the chapter, we will study the history and the evolution of those legislation and how it enhances our privacy.

By integrating data protection techniques like dynamic data masking, entities not only fortify their defenses against breaches but also align more closely with emerging legal standards, potentially reducing liabilities.

In the following sections, while we explore the complexity and aftermaths of data breaches, it's essential to remember that tools and strategies of data masking can serve as formidable allies in the quest for enhanced data security.

1.1 Understanding data breaches

Data security is one of the most important topic for a company or any organisation. When your data is compromised by an unauthorised actor, we call it a data breach.In the occurrence of a data breach, malicious actors want to access and steal sensitive information which can range from fiscal code, credit card details,passwords, personal data...

In this chapter we are going to go in depth about data breaches,the impact it has on a personal and global level, how to prevent yourself and your business from it and the regulatory aspect of data breaches.

Data breaches can occur in many different ways, an organisation must always be cautious and updated to know how to protect their data.

1.1.1 Why do data breaches happen?

According to the World Economic Forum of September 2022, human error causes around 95% of the security incidents [1]. There are many factors that make a human the weakest link in security such as poor password or not updating it more regularly. Passwords are not the only point of entry in data breach but users also can share information with unauthorized individuals.

Aside from the human error,there are many other ways a data breach can happen,the most common example would be a malware that steals confidential information or encrypts the victim's data.

1.2 Statistics on data breaches

Data breaches has a serious impact on businesses and people and here we are going to see it's consequences. According to 2022 IBM data breach report [2], the average cost of the data breach is 4.35 million dollars. This is a 2.6% increase from last year and more than 10% increase from 3 years ago, see Fig. 1.1.

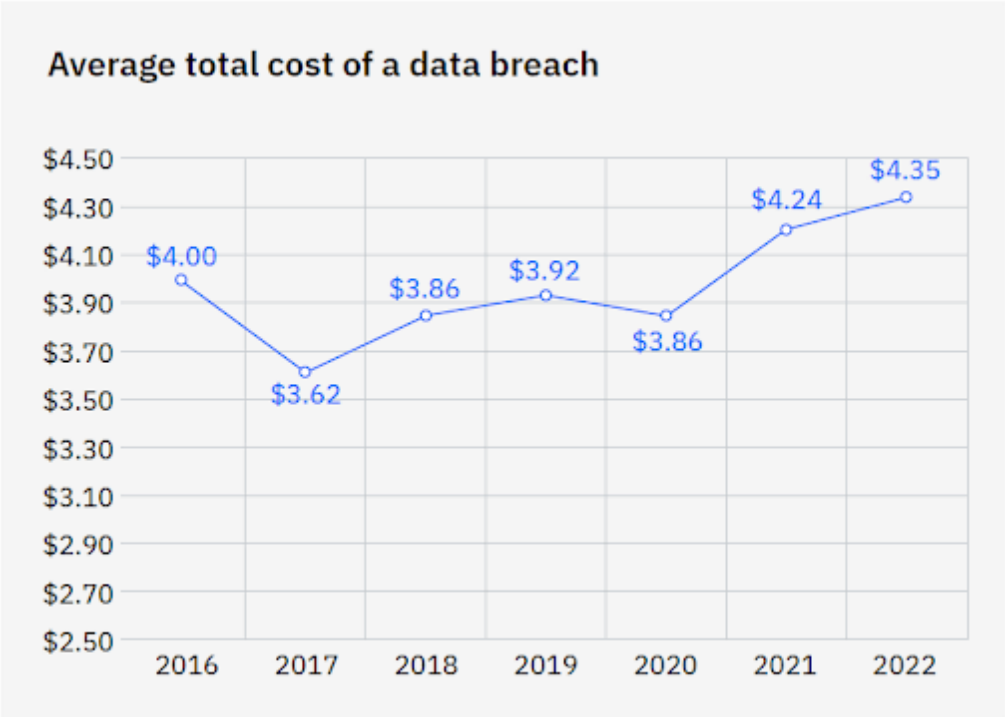


Figure 1.1: Cost of data Breach Report IBM 2022 (source: Bitsight).

Country wise, Italy is the 8th on the ranking list of countries affected by the data breach where the estimated amount of money lost is 3.74 million dollar from a value of 3.61 million dollars in the last year.

Research shows that data breach caused by poor security controls is responsible for billions of exposed records. During the year 2022, the exposed records were 19.81 billions according to the 2021 cyber Risk analytics breach review [3].

Most of these data breaches impacts 3 main sectors, the most affected industry is the healthcare industry where as of 2022, data breaches has cost 10 million dollars. Finance and technology industries follows up with a cost of 5 million dollars each [4].

1.3 Impact of data breaches

Entities affected by a data breach suffers heavily. The main impact is revenue loss. These losses can be caused by downtime payment and audit fees. It can even lead to the reputation tarnish which would force clients to find reliable business offering the same service. Not only the loss of clients might lead to a revenue loss but also legal fees and regulatory fines are among the expenses caused by a data breach.

1.4 Case studies of major data breaches

In this section we are going to focus on different sectors affected by data breaches. Some was from human errors and others were misconfigurations [5].

1.4.1 The First American Corporation

First American Financial Corporation is an American financial services company which provides title insurance and settlement services to the real estate and mortgage industries. As of May 2019, there was a data breach and compromised more than 885 million financial and personal records. The error was due to a web page not protected by an authentication process.

Among information obtained through this data leak, there was buyer's and seller's Names,ID,Social security Numbers, driver's licences, where they reside,Email Addresses and contact information. This is particularly dangerous as all these information could have been used for a phishing attack [6].

After this incident,The American Corporation was found to have breached cybersecurity protections regulations and The New York State Department of Financial Services fined them $487,616 [7].

1.4.2 The city of Calgary data breach

An employee from the city of Calgary has created a data breach when he unintentionally shared confidential information with another employee from a different municipality of the same province. The employee infringed a privacy violation sharing these information by email and the data was not encrypted.

This was not a malicious user though, as this leak was accidental. The employee needed technical advice and sent the data. This creates a serious risk as the data was not encrypted. The concerned individuals face a high possibility of identity theft and financial scam [8].

1.4.3 US voters records 2017

This unprecedented data breach affected a data science company called Deep Root located in the United States. The company affected played a key role in President Donald Trump's electoral campaign. This data breach affected up to 198 million US citizens.

The breach was discovered by a security specialist who encountered an unprotected database containing details of those US registered voters.

Inside the database were information such as "names, birth dates, residential contacts, telecommunication details, and voter enlistment specifics". Not only personal details, the database also showed predictive models suggesting voter behaviors, policy affinities, and potential candidate support levels.

Upguard emphasized the troubling fact that this goldmine of information was not shielded in any way, making it a potential target for "anyone with online access" [9].

1.4.4 Solar wind

SolarWind is an American software company. It develops and manages a range of system management tools including but not limited to network and infrastructure monitoring. Its products are used world wide by a big number of organizations and companies. One of their most popular product is called Orion. It is an IT performance monitoring system tool.

As Orion is an IT monitoring system, it has privileged access to IT systems that allows it to collect logs and system performance data. These advantages make it a more profitable target for cyberattacks.

To add salt to the wound, this product is considered to be used by more than 30,000 public and private entities, including agencies to monitor and oversee their IT assets.

The data breach occurred when SolarWinds introduced a backdoor malware when updating the Orion software. This gave unauthorized access of the user's systems to threat actors [10].

This incident compromised data, networks and system on a massive scale, affecting thousands of organizations.

The SolarWinds hack was a significant event, and its importance doesn't solely stem from the breach of one company. Instead, it had a much broader impact because it set off a chain of events that affected a vast network of organizations, including the U.S. government. This incident revealed vulnerabilities in the interconnected systems that many organizations rely on, emphasizing the need for enhanced cyber security measures and vigilance in protecting digital supply chains.

According to Solar Wind quarterly report of 2021 this data breach cost the company $40 million [11].

1.5 Legal aspects of data breaches

Data breaches often results out in legal case files. In this sub chapter we are going to analyse the history of data laws and some of the major data breach regulation that has changed the world of data protection either for the users or companies.

Finally we will analyse some of the historical penalties and compensations imposed by those regulations to companies that failed to comply with them.

1.5.1 History context of data breach laws

Data breach law is not a new concept, it dates back in the nineteen hundreds with the U.N declaration of Human rights. In this declaration, it's article 12 states that "No one shall be subjected to arbitrary interference with his privacy,family,home or correspondence, nor to attacks upon his honour and reputation.Everyone has the right to the protection of the law against such interference or attacks".

In 1974, The US department of education elected the Family Educational Rights and Privacy acts (FERPA). This federal law protects the records of the students in all schools, universities and institutions.

Finally in 1995, the European Union adopted the Data Protection Directive that control how the companies handle personal data of their users. This is a more restrictive law compared to its counterpart of the US.

It was replaced in 2018 by the recent one, the GDPR.

The Health Insurance Portability and Accountability Act (HIPAA) from 1996 aims to simplify healthcare information processes, safeguard personal data in healthcare, and address health insurance restrictions.

California introduced laws in 2003 that made businesses and state agencies tell people when their personal data was at risk(if there have been a breach). Many other places in the U.S. and around the world have since made similar rules based on California's example. This law was called a state data breach notification laws.

In 2018, the European union implemented the General Data Protection Regulation laws that protects the data and the privacy of in the European Union EU and the European Economic Area EEA. This law also goes into effect for the movement of data outside of EU and EEA.

In 2020, The state of California put in place a statute that controls the information and data of California residents and how businesses handle them. The statute was signed into a law and put into action on January 1,2020. The law was called The California Consumer Privacy Act(CCPA).

The CCPA created a precedence to other states to create their own statute regarding privacy and data handling. In 2021, Virginia signed the same Consumer Data Protection Act into a law. The same was also signed in 2023 by the Colorado state.

1.5.2 Penalties and compensation

We are going to analyse penalties and fines given in Europe imposed by the GDPR.

The GDPR gave European Data Oversight agencies the power to charge companies up to 4% of their annual earnings for data mismanagement. Before this, the fines were inconsistent: Spain's regulatory body could charge a maximum of €600,000, France's CNIL set a cap of €150,000 for initial violations or up to €300,000 for subsequent ones, and the UK's 1998 law had a limit of £500,000, though it was later updated in 2018 to match the GDPR. Post-GDPR, the response across countries was varied: some were quick to impose hefty penalties, while others acted more conservatively. Notably, during this time, Croatia, Estonia, and Slovenia didn't publicize

any GDPR-related fines [12].

The graph 1.2 shows how countries fines companies for data privacy non compliance or security violations [13].

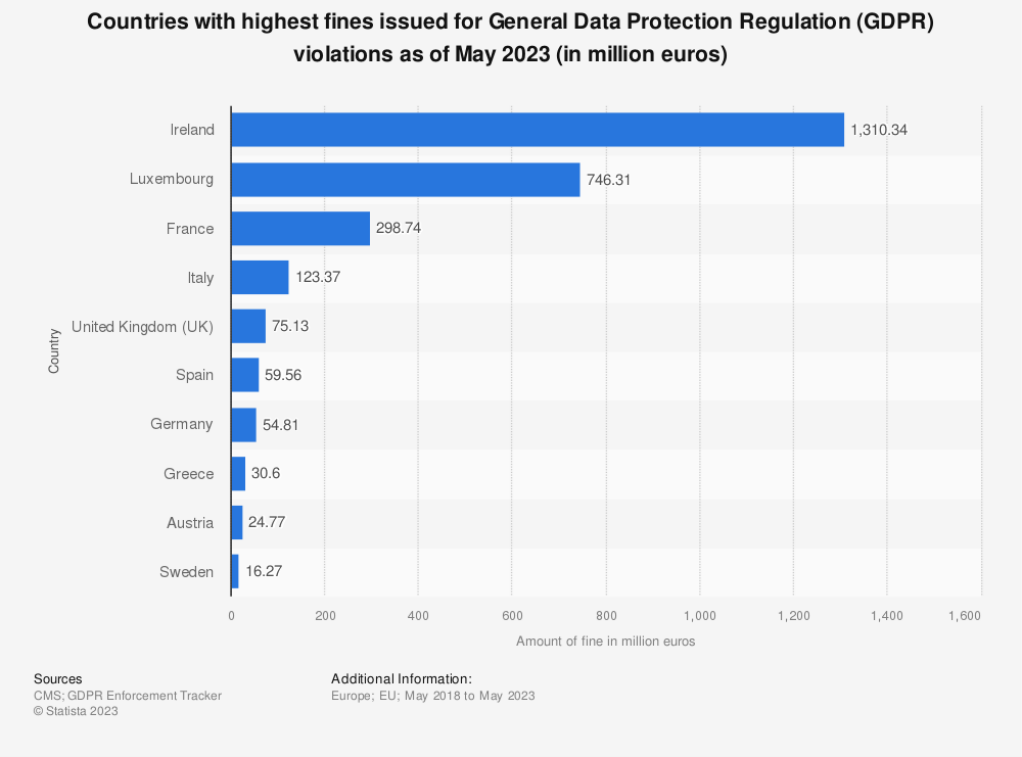


Figure 1.2: Fines issued for GDPR by country (source: Statista).

Ireland has a precedence in fining countries due to a historic €1.2 Billion on Meta (former Facebook). This record breaking fine was imposed due to an insufficient data protection process that was used when transferring personal data of European users to the united states [14].

This fine is not the first one imposed to Meta as of 2022, the Ireland's Data Protection Commission imposed another €405 million due to the fact that the processing of children's personal data was not in order with the legal bases. Meta is not the only big company that is facing fines due to irregularities and failure to comply. Tiktok also faces a fine of €345 million due to the malpractice of how to manage account datas of children.

As for Luxembourg, over 95% of the fine issued by the Luxembourg National Commission for Data Protection (CNPD) was imposed to Amazon in relation to its processing of personal data and compliance with data protection laws.

As for Italy the figure below shows which company has been fined the most and how much it was fined since the implementation of the regulations of data protection.

As shown in the figure 1.3 The Italian data protection(Garante) has imposed fines on a lot of big name technology companies in the world. Recently it fined a sum of $20 million to Clearview AI fine for non compliance. This company who owns a database of over 10 billion facial images worldwide had taken part in illegal surveillance activities with the country. Tech companies are not the only ones being targeted by this organization, as also the telecommunication company WIND had been issued by Garante a fine of €16 million. This fine was due to complaints from multiple individuals who received unsolicited marketing activities via calls and SMS.

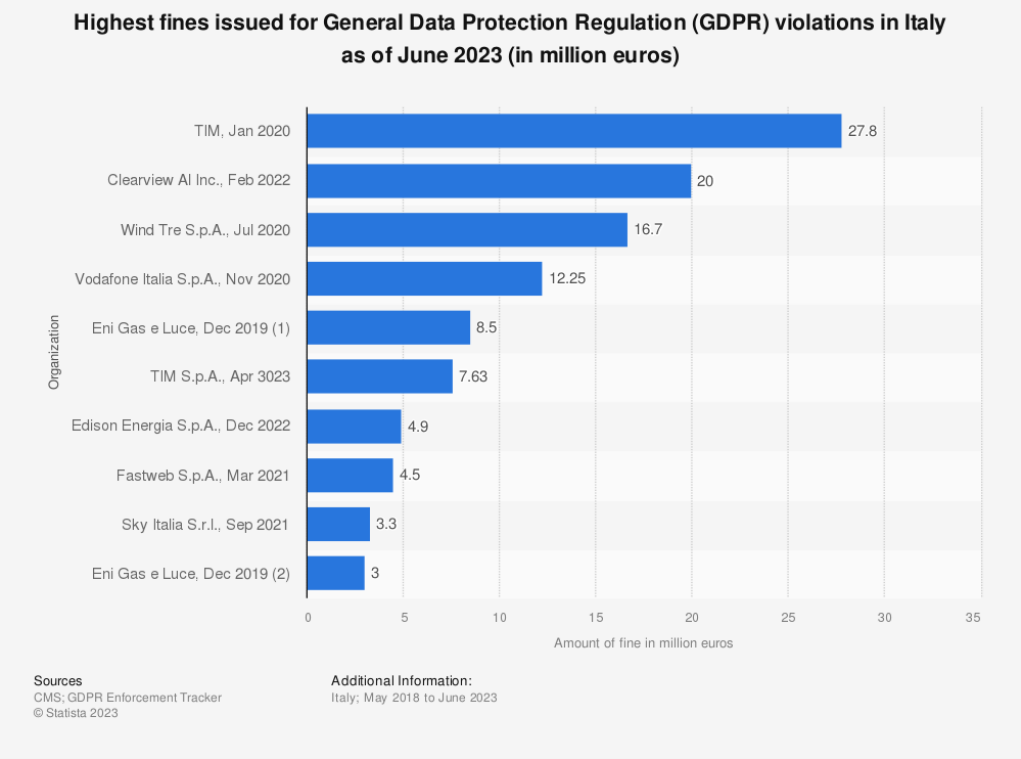


Figure 1.3: Italian historic Fines issued for GDPR (source: Statista).

1.6 Conclusion

In this chapter, we took a deep and detailed look on breaches, what they were, how they happen and its impact on both the company targeted, the clients and the society in general.

We studies major data breaches and what caused them and we took a look at the legal grounds of a data breach. Starting with the historical context up to statistics on penalties and compensation for both Europe and in Italy more specifically.

In legal actions we emphasized our study only on GDPR as it is the direction the thesis is going to take in later chapters and more importantly this thesis is the result of an internship held in the European Union where GDPR is the more prevalent data protection law.

CHAPTER 2

BACKGROUND

In this chapter we are going to a make a thoroughly comprehensive analysis of data masking and the research done so far on the subject. We will take a look at the principles and objectives along with the methods used so far to protect our data through masking and finally we will examine the best practices used when carrying out data masking.

2.1 The five laws of data masking

1. **Masking is irreversible** The original information we are masking can not be retrieved from the masked result.

2. **The result should accurately reflect the characteristics of the original data** One of the main reasons of using masking instead of just generating random values is that we need to have an outcome with similar features and attributes to the original values. An example of this law would be passing the masked result for testing developers to examine,verify and validate the program. You would need to have a sample of the data with similar aspects and not random values.

3. **Consistency in references must be preserved** In a database if a value is a primary key and it is masked in a certain way, we must make sure that in the other tables where that value is referenced as a foreign key should also be masked in order to maintain the consistency.

4. **Mask non-sensitive data only when there is a possibility of reconstructing sensitive information from it** It is not mandatory to mask every element in your database; only the portions you consider sensitive require masking. However, certain non-sensitive data might still have the potential to reconstruct or link back to sensitive information. For instance, if you obscure a customer ID but other transaction details uniquely correspond to a single customer, it's essential to also mask those details. This scenario,known as inference analysis, needs to be addressed by your data masking solution to ensure comprehensive protection.

5. **The masking procedure should be a reproducible and consistent process** Masking data just once is tough to maintain and doesn't work effectively.

It is crucial to make sure that development and test data closely resemble the rapidly changing real data. Without automation, masking is inefficient and expensive [15].

2.2 Methods of data masking

There are diverse masking methods that works in different ways depending on use cases and scenarios. Depending of the user's policies, capacity and requirements, the company may choose to implement one accordingly. we are going to analyse most of the popular methods of data masking and give a complete picture on why the company chooses to implement one instead of the other.

2.2.1 Static Data masking

Before discussing about static data masking, I would like to look at the word "Realism". This concept is highly important in masking as data created by masking process must be realistic. According to the oxford dictionary, Realism is a quality of representing something in a way that is accurate and true to life. We adopt this characteristics when creating our data as we need our masked data to have the same features as our real sensitive data. An example would be masking a string to a number in a development or testing environment. This would create a testing problem as we can not handle a number as we can handle a string. Realism is not only a concept for static data masking as also other types of masking would have the same issue if not handled cautiously. Static data masking is a technique of masking where sensitive data are replaced permanently at rest. This masking technique is used mostly when creating development and testing environment where real data is not necessary but fake realistic ones [16].

How to implement static data masking

As this masking is all about masking data at rest, we need to take the database to mask as input and apply transformation on sensitive data. While transforming sensitive data keep in mind all the foreign key that might be referential between databases. And lastly produce a realistic high quality data database with masked values.

The figure 2.1 explains graphically what we discussed above where in order to create a static data masking requires implementing the masking with the data at rest and creating a new database where those data would be stored.

Advantages of static data masking

- No risk run when the masked database is exposed since there is no sensitive data.
- Unlike the dynamic data masking there is no performance overhead on the masked database. In static data masking the obfuscation operation is done before the database is made available to the users. This means that there is no operation that lies between the user and the masked database that could hinder or delay the operation.

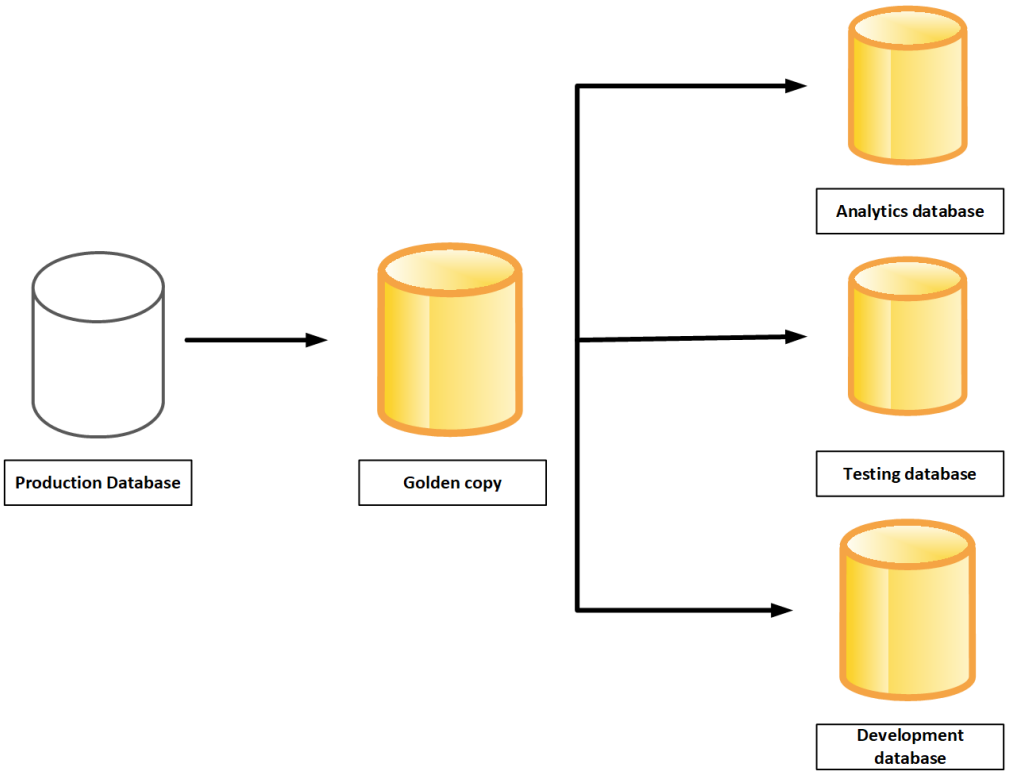


Figure 2.1: static data masking

- Very simple to implement as it does not need to write a detailed object level security as all sensitive data has been replaced [16].

Disadvantages of static data masking

- Since masking is not done in real time, it requires an ahead of time masking process which might hinder the business plans as the time it takes for completing the masking of database depends on the size of the data to mask.

- In some use cases, static data masking is not applicable as it modifies completely the database, we can not use it to protect the production environment.

2.2.2 Dynamic Data masking

Dynamic data masking which is the main topic of this thesis is when the masking instead of being implemented on data at rest, it is done while data is in-transit.

This means that dynamic data masking is designed to protect the data in real time. In order to have this type of masking, a proxy is required in order to handle the query and communicate with the database. The proxy lies between the user and the database system.

The advantage with dynamic data masking is that it is fine-grained and has the capacity to provide protection basing on the roles of the user who queried the database.

How to implement dynamic data masking

In order to make a successful DDM masking operation, you need to place a proxy running the dynamic data masking function between the database and the query. Lastly you would need to implement data masking policies the DDM Tool would base it's self to provide a detailed access and masking of the data to appropriate users.

There are 2 ways the proxy returns the masked result to the user. Usually the database query masks database by modifying the query sent to the database. The other way would be to execute the query AS IS from the user but modifying the result set returned by the database. All queries need to pass through the proxy before getting to the database.

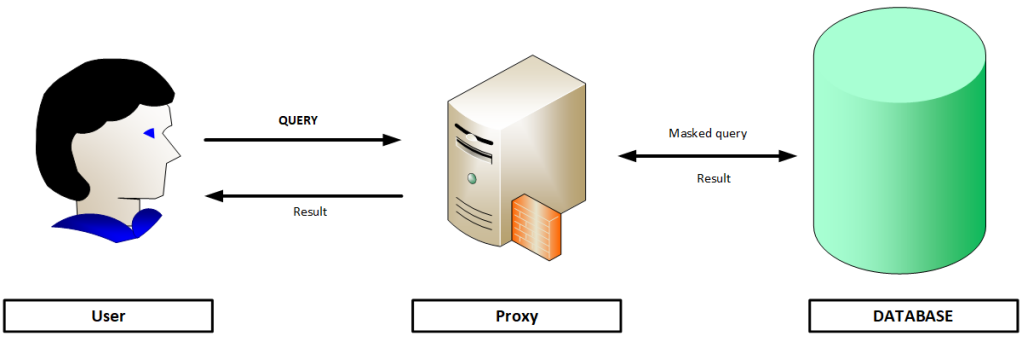


Figure 2.2: dynamic data masking

The figure 2.2 details how the operation involving querying the a database secured with a dynamic data masking proxy. As shown above, the user does not interact directly with the database but, he interacts with the proxy which can modify his query according to his role and the policy established. Once the query is reviewed by the proxy, it queries the database which will provide the data to the user through the query again. These data comes back to the user already protected.

Advantages of dynamic data masking

- An additional and customised layer of security added to protect the data
- Real time protection is provided which makes it suitable to protect the data in production.
- Unlike static data masking that requires a certain time ahead to mask the data before presenting it to the users, in DDM, there is no processing time needed in advance for masking as everything is done in real time [16].

Disadvantages of dynamic data masking

- Dynamic data masking protects the data only in read mode scenario, in case we need also to write on the same database, there might be conflict where we will write on the database masked data thus corrupting it.
- Compared to the static data masking there is a performance overhead as you do not connect directly to the database but you pass by the proxy.

- The proxy becomes a single point of failure as all the queries from the users to the database passes through it.The biggest issue with this is that if a user manages to evade using the proxy, he would get the data stored in the database with unprotected sensitive data.
- there is a lack of maturity for this technology as it is relatively unlike static data masking.

2.3 Common data masking methods

There are different methods to implement data masking, and depending on the type of data the company have to mask and what are the company's policy, we may choose among the following:

2.3.1 Shuffling

Similar to the substitution method the shuffling methods uses the same masking column and shuffles it in random positions.

An example would be shuffling the student's names columns across a database, and putting each name in a random position depending on the shuffling algorithm used. The result would be similar to the original data.

The shuffling method of masking is prone to an attack, as if a malicious user knows the algorithm used, he would be able to reverse engineer the algorithm and get the original database [17].

2.3.2 Substitution

This is the simplest form of masking where a value is simply substituted into another one of the same type. As en example, the masking might change the person's phone number with other random phone number [17].

2.3.3 Scrambling

Scrambling is a data masking technique where we change the order of characters in the original data. This type of masking can only work on some types of data and not others. Due to its simplicity, scrambling technique is not very secure as someone who remembers the order of the original data can immediately be able to decipher the scrambled values.

An example would be changing the original employee ID in the production environment from ID: 67598 to ID:89576 [18].

2.3.4 Data variance

The Number Variance method is an effective strategy for altering numerical or date-related data. This approach modifies each value in a dataset by adding or subtracting a random percentage of the value's true amount. The primary benefit of this method is its ability to obscure the actual data while maintaining the overall scope and distribution of the dataset within its original parameters. For instance, in a dataset

containing a debt figures, each value can be customized and changed by a factor of 10 percent.

Consequently, the altered figures will remain relatively close to their true values. Similarly, this method is applicable to date data, such as birth dates, which can be randomly adjusted within a specified range [17].

2.3.5 Nulling out

Nulling out is another type of data masking where data is nulled independently of which type of data. The advantage of this method is that, it is very simple and straightforward; however it has also a big drawback as data integrity is no longer present.

This makes it especially difficult if the masked data is used for development and testing purposes as it is no longer possible to use it to different use cases possible with real data [17].

2.3.6 Masking out data

This is a generic term that means anonymisation but it is also a masking method where data is in certain fields is replaced by a standard character usually 'X'. This perfectly masks our data at the same type keeping the format of the data as it keeps the properties it had before.

An example would be a credit card with numbers: 2234-7146-3088-6173. This same credit card number becomes 2234-XXXX-XXXX-6173.

As explained above, When we replace raw data with a character, we remove the sensitivity of the content and still preserve the format of our data, we have to be careful to know how many characters we anonymise as also which ones to mask. This is due to the fact that, if you remove the first characters in the credit card, u will strip the possibility to recognize which type of credit card it is. In few words, the masking of credit card as The masking characters effectively remove much of the sensitive content from the record while still preserving the look and feel. It would not be hard to regenerate the original credit card number from a masking operation such as: XXXX-XXXX-XXXX-6173 would not be suitable.

Same with the masking of the last few numbers as 2234-7146-3088-61XX would be too easy to revert since credit cards use a fairly public algorithm for checksum.

In conclusion, the decision process of making data anonymous should consider all the above risks [15].

2.4 Applications and Use Cases of Data Masking

Data masking has different applications, but most companies go through this process with their data for security reasons.

Legal Obligations: the regulations in which companies have to be compliant are becoming more and more broader and specific. It's widely anticipated that the criteria for data security and management will only tighten further.

Loss of confidence in the client trust: in many regions, the repercussions of a data breach extend beyond regulatory penalties. Such incidents might not be

the primary concern immediately. The fallout from data breaches, intentional or accidental, can be severe.

Consider the repercussions for an organization if potential clients hesitate to share sensitive data due to negative media coverage about a data breach. The costs involved in managing the brand's image and addressing public concerns can significantly surpass any legal fines. The impact on senior management, who may have to engage in damage control and reassure stakeholders, can be profound. The financial implications of tarnishing a company's reputation often overshadow any regulatory fines.

Unintentional Disclosure: the potential for inadvertent data leaks is sometimes overlooked in the context of security risks related to actual test data. The assumption that masking test data is unnecessary because everyone has access to the production environment is misguided.

The threat of accidental data leaks persists. Masking critical data elements (like credit card numbers or customer email addresses) can help reduce the risk of damaging leaks while maintaining the operational functionality of the databases.

Data utilization: data obfuscation enables organizations to utilize representative data for the purposes of testing and development. By employing this technique, genuine sensitive data remains concealed, ensuring that the integrity and confidentiality of the information are upheld.

This practice is crucial for verifying the proper functioning of software and systems, without compromising data security [19].

Secure third-party data sharing: In situations where companies are required to distribute data externally, data masking acts as a safeguard, concealing confidential elements within the data.

This process facilitates secure information exchange and fosters collaboration between business partners, all while maintaining data privacy [19].

Cost-efficiency: The repercussions of data breaches can be extensive, including legal penalties, damage to reputation, and the costs of rectification.

Implementing data masking can mitigate the risk of such breaches, offering organizations a protective measure that can lead to significant savings in terms of both time and financial resources [19].

2.5 Masking Constraints

As said above, data masking hides the data but keeps the integrity of the data. However, this process can encounter several challenges, particularly when specific data properties must be preserved. These constraints make the development of masking algorithms complex, necessitating sophisticated solutions to balance data utility with privacy [20]. Below are key constraints identified in data masking and suggestions for enhancing clarity:

- Format preservation: the masked data must retain the same format as the original data, meaning the length and structure of the data should remain consistent. For instance, a 16-character codice fiscale must be replaced with a masked equivalent of the same length and alphanumeric composition.

- Data Type preservation: in environments like relational databases, it's crucial to maintain the data types across masked values [15]. This ensures compatibility with the database schema, preventing issues like inserting textual data into fields designated for numbers or dates.
- Gender preservation: attributes tied to gender should be masked with alternatives reflecting the same gender. This is particularly relevant for names, where a female name should be replaced with another female name, and similarly for male names.
- Semantic integrity: masking must not only preserve the format and data type but also ensure that the data remains valid under specific checks, such as a credit card number's validity criteria.
- Referential integrity: the relationships between elements across different tables or files must be meticulously maintained in the masking process to preserve data relational context and meaning.
- Aggregate Value: the overall statistical properties, such as totals or averages, of a masked dataset should closely or exactly mirror those of the original data, ensuring analytical value is not compromised.
- Frequency distribution: the masking process should accommodate the need for either random or logically consistent frequency distributions. For example, maintaining geographical relevance in datasets by preserving or logically altering pin codes to reflect realistic geographic distributions.
- Uniqueness: ensuring the uniqueness of masked values is paramount, especially for identifiers or keys that must remain unique to maintain referential integrity across database tables.

By addressing these constraints, data masking strategies can effectively secure sensitive information while preserving the utility and integrity of the masked data for various applications.

2.6 Best practices in implementing data masking

Initiate the data masking process by first pinpointing:

- Where the sensitive data resides.
- Who has the clearance to view this data.
- The specific use cases for the data.

It's unnecessary to mask every data element within an organization. Prioritize the precise identification of sensitive data in all environments, both production and testing. This identification process could be time-intensive, depending on data complexity and the company's structure.

Adopting a one-size-fits-all approach to data masking tools isn't feasible for larger organizations due to the varied nature of data. The selection of masking techniques

should be influenced by adherence to internal security standards, budget limitations, and may even involve developing bespoke masking solutions. Consistency in the use of these techniques is vital to ensure referential integrity across similar data types.

Protecting the methods and tools used for data masking is as critical as safeguarding the data itself. For example, the risk associated with a substitution method that uses a lookup file is significant if unauthorized parties access it. Implement strict access controls to safeguard the masking algorithms.

Transform data masking into a scalable, efficient, and automated routine. This adaptability prevents the need for redevelopment with every organizational, project, or product data alteration.

In conclusion, a robust data masking initiative should be holistic, including the identification of sensitive data, selecting and applying the correct masking techniques, and conducting ongoing audits to ensure the effectiveness of the masking efforts.

2.7 conclusion

In this section we explain in details what was data masking and the difference between dynamic and static masking. We evaluated also the different masking methods including shuffling,substitution and scrambling etc. Finally we showed masking constraints and how the masking algorithms would be different one another depending on the constraints followed. In the following chapter we are going to look into the different legal compliance laws around the globe and how it touches the subject of data security.

CHAPTER 3

DIFFERENT DATA PRIVACY AND SECURITY LAW

In the digital age, where information serves as a cornerstone of both opportunity and risk, regulatory bodies worldwide have enacted privacy and security laws to safeguard user data. This chapter aims to dissect the different compliance regulations, delving into specific articles that delineate the mechanisms and obligations designed to protect user data.

3.1 General Data Protection Regulation (GDPR) - European Union

The General Data Protection Regulation (GDPR), which took effect on May 25, 2018, is a significant piece of data protection legislation established to govern the processing, storage, and management of personal data from individuals residing in the European Union (EU, 2016). This legislation marks a substantial enhancement of the EU's data protection capabilities, aiming to address the privacy issues emerging from the rapid advancement of digital technologies. While GDPR is designed to safeguard the data of EU citizens, its influence extends globally, impacting any organization that engages with the European market or manages personally identifiable information of individuals within the EU [21].

3.1.1 Data security

The GDPR requires an appropriate and organizational measure to handle data securely. This is emphasized in the recital 78 citing [22]: "The protection of the rights and freedoms of natural persons concerning the processing of personal data require that appropriate technical and organizational measures be taken to ensure that the requirements of this Regulation are met. To be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures that meet in particular the principles of data protection by design and data protection by default. Such measures could consist, inter alia, of minimizing the processing of personal data, pseudonymizing personal data as soon as possible, transparency about the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and

improve security features. When developing, designing, selecting, and using applications, services, and products that are based on the processing of personal data or process personal data to fulfill their task, producers of the products, services, and applications should be encouraged to take into account the right to data protection when developing and designing such products, services, and applications and, with due regard to the state of the art, to make sure that controllers and processors can fulfill their data protection obligations. The principle of data protection by design and by default should also be taken into consideration in the context of public tenders" [23].

Technical measures encompass various actions, ranging from mandating employees to implement an authentication and authorization mechanism on accounts containing personal data [22]. Organizational measures involve initiatives such as conducting staff training sessions, integrating a data privacy policy into the company's employee handbook, or restricting access to personal data solely to relevant employees within the organization,

3.1.2 Data protection by design and by default

From now on, everything done in an organization must, by design and by default, consider data protection. Practically speaking, this means you must consider the data protection principles in the design of any new product or activity. The GDPR covers this principle in Article 25.

3.2 Health Insurance Portability and Accountability Act (HIPAA) - United States

The 1996 Health Insurance Portability and Accountability Act (HIPAA) established federal guidelines for safeguarding the private health information of patients, prohibiting its disclosure without patient authorization or knowledge. The U.S. Department of Health and Human Services (HHS) formulated the HIPAA Privacy Rule to apply these requirements. In addition, the HIPAA Security Rule was introduced to protect a specific segment of data protected under the Privacy Rule [24].

3.2.1 Who is covered by the rule

As the HIIPA act safeguards patient information, every organization or entity that collects or has any interaction with this type of information is under this compliance law. This includes health plans, health care clearinghouses, and any health care provider who transmits health information in electronic form [25].

3.2.2 What information is protected by the law

Not all user data are under HIIPA protection, the patient's data under the said law are only protected health information[25].

Individually identifiable health information encompasses information, along with demographic details, that pertains to:

- the individual's health status or medical conditions in the past, present, or future.
- any health care services provided to the individual,
- any past, current, or anticipated payments for health care services received by the individual and can either directly identify the person or there exists a reasonable possibility that it could be used for identification. This type of health information typically includes various common identifiers such as name, address, date of birth, and Social Security Number.

3.2.3 What would be the relationship between data masking and the HIIPA law

As we are protecting a user's data, the process of de-identification of information can be implemented by data masking. De-identification is a process in which data can no longer be associated with a certain user as all reference to him has been altered or rendered anonymous. These types of information are reported by the article 45 C.F.R. § 164.514(b). They include but not limited to:

- (A) Names;
- (B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
  - (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
  - (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
- (C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
- (D) Telephone numbers;
- (E) Fax numbers;
- (F) Electronic mail addresses;
- (G) Social Security numbers;
- (H) Medical record numbers;
- (I) Health plan beneficiary numbers;
- (J) Account numbers;

- (K) Certificate/license numbers;
- (L) Vehicle identifiers and serial numbers, including license plate numbers;
- (M) Device identifiers and serial numbers;
- (N) Web Universal Resource Locators (URLs);
- (O) Internet Protocol (IP) address numbers;
- (P) Biometric identifiers, including finger and voice prints;
- (Q) Full face photographic images and any comparable images;

During the process of de-identifying these data, a qualified expert uses a formal process in our case it would be masking, and concludes, through the application of statistical or scientific principles (use case testing), that the risk is very small that the information could be used to identify the individual.

3.3 California Consumer Privacy Act (CCPA) - California, United States

California set a precedent as the first U.S. state to enact a broad consumer privacy statute when the California Consumer Privacy Act (CCPA) took effect on January 1, 2020.

This law provides California residents with unprecedented rights concerning their data and places numerous responsibilities on specific businesses operating within California regarding data protection [26].

3.3.1 What is protected by CCPA

The legislation applies to the Personal Information of all individuals identified as California Residents. According to the Act, a "resident" is defined as (1) any person present in the State for reasons beyond a temporary or short-term visit, and (2) any person whose home is in the State but is currently outside the State for a brief or temporary reason [27].

3.3.2 Who must comply with the law

The CCPA applies to for-profit entities that gather and manage the personal information of California residents, operate within California, and satisfy at least one of the following criteria:

- Generate annual gross revenues greater than $25 million
- Annually collect, share, or receive the personal information of 50,000 or more California residents, households, or devices;
- Earn 50% or more of their yearly income from the sale of personal information belonging to California residents.

3.3.3 What qualifies the personal information

Personal information as defined by the CCPA is "information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household" [28].

Most personal information is almost the same as the ones seen in other regulations above and may include;

- Personal information is information that identifies, relates to, or could reasonably be linked with you or your household. For example, it could include your name, social security number, email address, records of products purchased, internet browsing history, geolocation data, fingerprints, and inferences from other personal information that could create a profile about your preferences and characteristics.
- Sensitive personal information is a specific subset of personal information that includes certain government identifiers (such as social security numbers); an account log-in, financial account, debit card, or credit card number with any required security code, password, or credentials allowing access to an account; precise geolocation; contents of mail, email, and text messages; genetic data; biometric information processed to identify a consumer; information concerning a consumer's health, sex life, or sexual orientation; or information about racial or ethnic origin, religious or philosophical beliefs, or union membership. Consumers have the right to also limit a business's use and disclosure of their sensitive personal information.

3.3.4 Correlation between CCPA with data masking

While the CCPA does not explicitly discuss de-identification or anonymization techniques outside of the definitions, it is an implicit strategy that is used to go around and sell information as no rule protects de-identified data.

3.4 Personal Information Protection and Electronic Documents Act (PIPEDA)- Canada

PIPEDA is a Canadian privacy law that regulates Organizations so that they obtain an individual's consent when they collect and use an individual's personal information.

3.4.1 How the act applies

Organizations that fall under the PIPEDA regulations are the ones who are in the private sector; collect use, and disclose personal information, and lastly use the information retrieved for commercial or for-profit activities in Canada [29].

3.4.2 What is protected by the law

The first critical step in adhering to PIPEDA is to grasp the nature of personal information as defined by the act, to identify the range of information your organization

manages, and to evaluate how these aspects intersect. Under PIPEDA, personal information encompasses data concerning identifiable individuals. This scope includes both objective and subjective data, applicable whether the information is documented or not.

The Office of the Privacy Commissioner of Canada (OPC) provides examples to illustrate what constitutes personal information, highlighting its broad spectrum:

Details such as:

- an individual's age, name, and identification numbers.
- An individual's ethnic background and blood type.
- Personal opinions held by an individual.
- Financial information including income levels, credit reports, and records of loans.
- Health-related information, encapsulating medical histories.
- The presence of any disagreements between a consumer and a business entity.

3.4.3 Correlation between PIPEDA and Data masking

According to Article 4.7 safeguards that says "The security safeguards shall protect personal information against loss or theft, as well as unauthorized access, disclosure, copying, use, or modification. Organizations shall protect personal information regardless of the format in which it is held. [30]"

As data can be protected differently we can also use dynamic data masking to give access to real data to verified users and mask personally identifiable information to unauthorized users.

In our case, data masking can also be used to de-identify information and use it for commercial reasons without falling under the compliance of PIPEDA.

3.5 The Data Protection Act (DPA) 2018 - United Kingdom

The Data Protection Act 2018 governs the handling of 'personal data', defined as information of individual persons. It grants individuals the authority to request access to their data via subject access requests and establishes specific regulations that must be adhered to during the processing of personal data [31].

3.5.1 Correlation between DPA and data masking

The DPA does not explicitly mention data masking, but the practice is relevant under the DPA's principles of processing personal data lawfully.

As DPA also bases itself on GDPR, there is a reference to Article 32 of the GDPR, titled "Security of processing", which requires Data Controllers and Data Processors to implement technical measures that ensure a certain level of data security appropriate for the level of risk presented by processing personal data [32].

With this in mind, data masking can be implemented as a way of securing user data or de-identifying it in a way that it will not fall in the GDPR domain.

CHAPTER 4

DATA MASKING TOOLS

The field of data masking tools is broad, encompassing varied solutions that service its many requirements: from ensuring compliance with regulations such as GDPR and HIPAA to fostering a secure environment for software testing. This chapter will take us through the key features in an attempt to understand what to look out for and choose the right tool that will meet your needs.

4.1 Features to Look For in a Data Masking Tool

As data masking can be used for different purposes, it is difficult to have a unique general tool that satisfies all the requirements. When evaluating data masking tools, several key features and capabilities should be considered to ensure they meet your security and usability requirements [33]:

- **Data format**: Various masking tools are equipped to manage distinct data sources and formats, including relational databases, flat files, XML and JSON files, as well as data stored in cloud services. Additionally, certain tools are capable of masking data dynamically as it moves through streams, web services, or APIs. It's crucial to select a tool that can effectively mask your specific data types and sources, ensuring that the data's quality and integrity remain intact.
- **Masking techniques used**: Different data masking solutions employ a variety of methods and capabilities for concealing your data. These methods include, but are not limited to, encryption, hashing, substitution, scrambling, shuffling, and blurring. Moreover, certain solutions are engineered to preserve the data's referential, functional, or statistical integrity. It is crucial to select a solution that provides the necessary masking techniques and attributes that meet your specific requirements for data analysis.
- **Performance and scalability**: The performance and scalability of your data analysis can be affected differently by various masking tools. For instance, certain tools offer the capability to mask data in real-time, upon request, or through batch processing. Additionally, some tools are designed to operate in parallel, distributed, or cloud-based settings efficiently. It's essential to choose a tool that can mask your data with minimal impact on speed or interruption to your data analysis workflow.

- **Ease of use and ability to customize**: It varies significantly among data masking tools, impacting their suitability for your data analysis projects. For instance, certain tools may offer intuitive interfaces, templates, or guided wizards that simplify the setup and execution of data masking tasks. On the other hand, some tools provide more sophisticated features, such as customizable scripts or APIs, enabling you to tailor and automate masking processes. It's important to select a tool that aligns with your technical expertise and project requirements, ensuring it complements your workflow and skill set effectively.

- **Security and Compliance features**: The security and compliance capabilities of data masking tools can significantly differ, impacting their effectiveness in meeting your data protection objectives. For instance, various tools might incorporate features like encryption, authentication, authorization, along with auditing and logging to enhance the security and traceability of your masked data. Moreover, some tools are designed with compliance in mind, offering built-in rules, policies, or benchmarks to assist in conforming to specific regulatory or sector-based data protection standards. Identifying a tool that robustly defends your masked data against unauthorized use or access is essential for ensuring data safety and regulatory compliance.

- **Price**: The pricing and support offerings of data masking tools vary, affecting their alignment with your budgetary constraints and technical requirements for data analysis. Tools may range from being free and open-source to licensed with diverse pricing strategies, including per-user, per-project, based on the volume of data, or specific features. Additionally, support services differ among tools, encompassing a spectrum from comprehensive documentation and tutorials to online forums and dedicated customer support. It is crucial to select a tool that not only fits within your financial parameters but also provides the necessary technical support to achieve your data analysis objectives.

Selecting an appropriate data masking tool requires a detailed evaluation of these characteristics, taking into account your unique requirements for data security, the regulatory landscape, and your operational framework.

4.2 Different masking tools on the market

Different data masking tools are available on the market as it is a booming market and they meet different needs. Here we are going to look into some of the most relevant masking tools on the market as of this year.

4.2.1 Delphix

The Delphix Dynamic Data Platform[1] offers an automated solution for securing non-production environments by substituting sensitive details such as social security numbers, patient records, and credit card information with realistic, yet fictional data.

[1] `https://www.delphix.com/solutions/data-compliance-security`

Contrary to encryption methods, which can be circumvented by acquiring user credentials, masking ensures irreversible data protection in downstream environments. Delphix's ability to uniformly mask data while preserving referential integrity across diverse data sources gives it an edge over competing solutions, all without necessitating any programming skills. Additionally, the Delphix platform integrates data masking with data delivery features, safeguarding sensitive information prior to its use in development and testing, or before being transferred to offsite data centers or the cloud.

As a web-based application accessible to multiple users, Delphix Masking delivers a comprehensive, secure, and scalable solution for discovering and masking sensitive data, thereby satisfying the demands of enterprise-grade infrastructure [34].

Delphix has several key features and characteristics that enables it to mask sensitive data in a company.

- **End-to-End Masking**: The Delphix platform automatically identifies sensitive information, applies irreversible data masking, and then produces reports and sends email alerts to verify that all confidential data has been securely masked.
- **Realistic Data**: Data masked using the Delphix platform maintains a production-level quality. Even in non-production settings, the application data that has been masked retains its full functionality and realism, facilitating the creation of superior-quality code.
- **Referential Integrity**: Delphix ensures uniform masking across diverse data sources by scanning both metadata and data to recognize and maintain the primary/foreign key relationships among elements. This approach guarantees consistent data masking across various tables and databases.
- **Algorithms/Frameworks**: Delphix framework comes equipped with more than twenty-five ready-made algorithms designed for masking a wide range of data types, from personal identifiers like names and addresses to sensitive information such as credit card numbers and textual content. Additionally, Delphix includes predefined profiling sets tailored for managing healthcare and financial data.
- **Ease of Use**: Delphix offers a unified solution that enables its customers to mask data across multiple platforms effortlessly. The platform's web-based interface allows for easy data masking through a simple series of mouse clicks, requiring minimal training for users.
- **Automated discovery of sensitive data:** The Delphix Profiler simplify the process by automatically detecting sensitive information within databases and files, significantly cutting down the labor-intensive tasks typically associated with a data masking project.

Delphix platform architecture

The Delphix Dynamic Data Platform comprises four key services, each crucial in providing updated, secure data to all who require it. These includes:

- **Virtualize**
- **Identify and Secure**
- **Manage**
- **Self Service**

The image below explains the architecture workflow of this product [34].

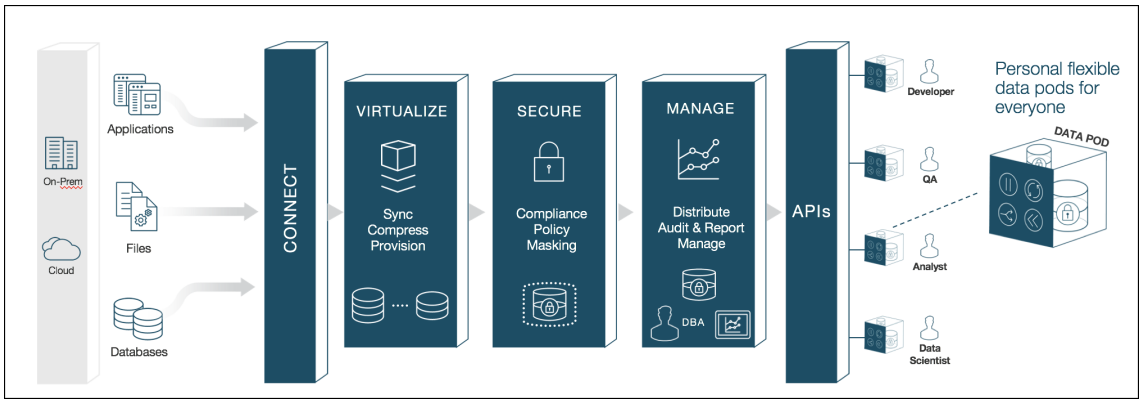


Figure 4.1: Delphix architecture (source: MaskingDocs).

Virualize
Delphix efficiently compresses collected data, typically reducing it to a third or even less of its original volume. Leveraging this minimized data footprint, Delphix then virtualizes the data, enabling the creation of streamlined, virtual copies. These copies are entirely independent and support full read/write capabilities. They can be rapidly deployed or decommissioned within minutes. Remarkably, they occupy only a small fraction of the storage required for traditional physical copies, with the capacity to accommodate ten virtual copies within the storage space that would normally be occupied by a single physical copy

Identify and Secure
The Delphix platform consistently safeguards private data through its built-in data masking feature. It protects sensitive information, such as names, email addresses, patient records, and social security numbers, by substituting these details with fictitious but plausible alternatives. Delphix is designed to automatically detect sensitive information and then enforce either custom or pre-established masking algorithms. By merging data masking with the data provisioning process within one unified platform, Delphix guarantees a streamlined and reproducible approach to delivering data securely.

Manage
Data operators are now able to swiftly provide users with secure copies of data in their designated environments, completing the process in just minutes. The Delphix platform acts as a centralized control hub for managing these copies. It gives data operators complete oversight and control over downstream environments, allowing them to efficiently conduct audits, monitoring, and reporting on access and utilization.

Self-service

Gives various users, including developers, testers, analysts, data scientists, among others, the flexibility to adjust data as needed. Users have the ability to update data to mirror the most recent production status, revert environments back to an earlier state, save data copies for future reference, create branches of data copies for simultaneous work on different versions, or conveniently distribute data among colleagues.

4.2.2 K2View

K2View[2] is an ideal choice for safeguarding a substantial amount of sensitive data. This tool adopts a data product methodology that simplifies the implementation process and cuts down on both time and expenses, all while addressing complexities found in enterprise environments.

It leverages automatic discovery and data cataloging capabilities to identify, classify, and organize sensitive information. Additionally, K2View enables detailed searches at the level of database files and metadata. The tool offers the flexibility to use numerous pre-configured masking functions, including but not limited to, substitution, randomization, shuffling, scrambling, data type conversion, clearing data, and concealing parts of the data, ensuring comprehensive data protection [35].

Architecture overview

The K2View Data Masking Solution is structured around two primary components [36]:

- The LUDB CONFIGURATION
- The EXECUTION SERVER(S)

The LUDB Configuration is a detailed, version-controlled setup encompassing all aspects necessary for the data masking implementation, including:

- Parameters for connecting to source and destination systems
- Definitions of masking rules (further elaborated in the DATA MASKING FEATURES section)
- Rules for Extract, Transform, and Load (ETL) processes, which may cover data enhancement, validation, reporting, or checks for data integrity. For an in-depth understanding of K2View's ETL capabilities, refer to our Data Migration documentation.

The execution server(s) consist of a network of servers that are managed through the LU Studio. This setup is responsible for extracting, masking, transforming, and transferring data from its source to the designated target. Each server runs several threads of the ETL + MASKING ENGINE, facilitating the distributed execution of data masking tasks, which results in exceptionally high performance.

[2] `https://www.k2view.com/solutions/data-masking-tools/`

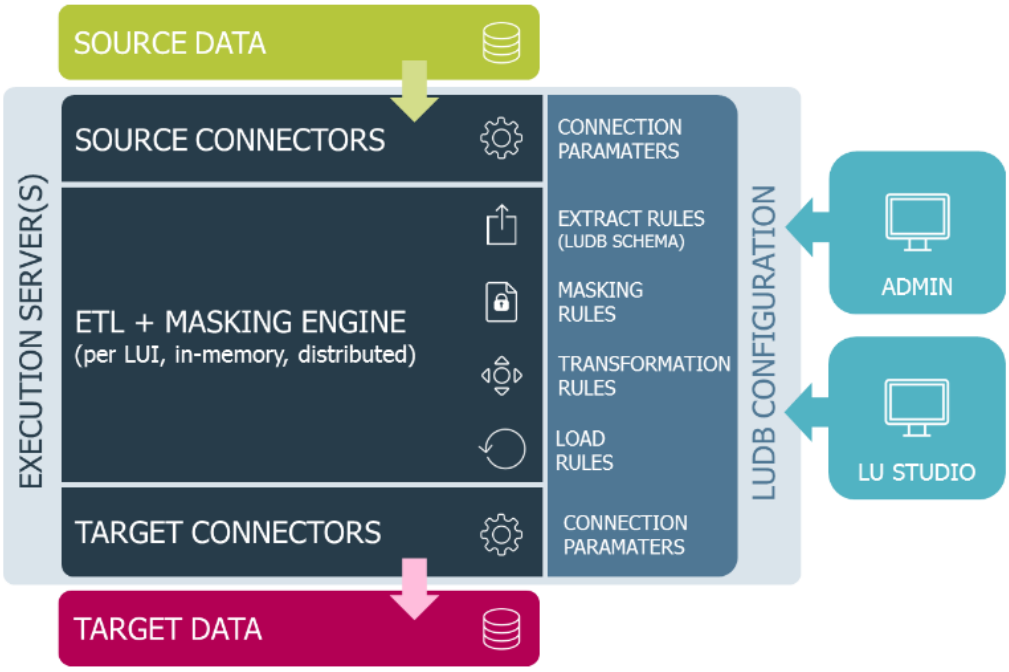


Figure 4.2: k2View architecture (source: MaskingDocs).

K2View Data masking features

- **End to End masking**: The Delphix platform seamlessly identifies sensitive information, applies irreversible masking to the data, and subsequently produces reports and sends out email alerts to verify the complete masking of all confidential data.
- **Realistic data**: Data masked by the Delphix platform maintains a quality akin to that of production data. Even within non-production settings, the application data that has been masked continues to be entirely operable and realistic, facilitating the creation of code of superior quality.
- **Referential Integrity**: Delphix ensures uniform masking across diverse data sources by analyzing both metadata and data to recognize and maintain the relationships between primary and foreign keys among elements. This approach guarantees consistent data masking throughout various tables and databases.
- **Ease of Use**: Delphix offers a unified solution that enables its users to apply data masking across multiple platforms effortlessly. Additionally, it eliminates the need for companies to develop custom masking algorithms or depend heavily on administrative support. The platform's web-based interface facilitates easy data masking with just a few clicks, requiring minimal training.
- **Automated discovery of sensitive data**: The Delphix Profiler streamlines the process by automatically detecting sensitive information within databases and files, considerably cutting down the labor-intensive efforts typically required for a data masking project.

4.2.3 Informatica

Informatica's[3] Dynamic Data Masking masks or restricts access to sensitive data depending on a user's role, geographic location, and permissions. It can notify administrators of attempts to access this data without authorization and maintains records for compliance and auditing purposes [37].

Informatica's Key benefits

- **Policy-Driven, Role-Based, Real-Time Data Protection**: Informatica Dynamic Data Masking implements real-time data privacy and security measures by dynamically masking, concealing, blocking, auditing, and alerting on unauthorized access. It offers customizable restrictions on access at various levels, including screens, tables, columns, rows, and individual cells.

- **Scalable and Easy to Install and Configure**: The software is capable of scaling up to support hundreds of databases through a single installation, facilitating rapid and uniform restriction of access across various tools, applications, and environments. This is achieved by establishing data masking policies just once and then applying them repeatedly. Moreover, these data-masking algorithms can be applied to sensitive data in any format.

- **Versatile and Nonintrusive to Applications or Databases**: Dynamic Data Masking is compatible with a variety of environments, including virtualized, traditional, big data, and cloud computing setups. It safeguards against unauthorized access to both custom and packaged applications, data warehouses, and operational data stores, all without affecting performance. Importantly, using Dynamic Data Masking does not necessitate modifications to existing applications or databases.

- **Integration With Authentication Software:** Informatica Dynamic Data Masking restricts access to sensitive business information, ensuring that only individuals who fulfill selected security criteria can view it. Additionally, the software integrates with your current identity management systems to hasten deployment and enhance the security coverage of your applications and tools.

- **Real-Time Data Masking and Blocking :** The software preserves the original data untouched and ensures that the functional look and consistency of the masked data remain intact for complex applications. It achieves this by synchronizing data values across different rows and tables.

Informatica's data masking Key Benefits

- **Cost-Effectively Protect Personal and Sensitive Data:** Informatica Dynamic Data Masking offers an economical solution for safeguarding your data against both internal and external threats of data breaches, without affecting the application's performance. It is straightforward to set up and instantly anonymizes sensitive data in real time.

[3] `https://www.informatica.com/it/products/data-security/data-masking/dynamic-data-masking.html`

- **Quickly Customize Privacy and Security Solutions for Maximum ROI:** Swift and simple to deploy across a variety of systems, including applications, backups, clones, data warehouses, as well as tools for development and database administration.

- **Support Digital Transformation, Privacy, and Cloud:** By ensuring secure entry to both production and non-production environments, your IT department can enhance efficiency and swiftly adapt to business demands. This support accelerates digital transformation, privacy projects, and the transition to big data and cloud-based assets.

4.2.4 Immuta

The Immuta Data Security Platform[4] helps organizations secure their data by providing sensitive data discovery, security and access control, and activity monitoring [38].

Immuta's features

- Sensitive data discovery
- Security and access control
- Continuous monitoring

Sensitive data discovery
Immuta automatically identifies and categorizes sensitive information, such as PII and PHI, within data lakes and warehouses, aligning with regulations like GDPR, HIPAA, PCI, among others. This process provides complete insight into which data needs access restrictions and ongoing surveillance [39].
Security and acccess control
The Immuta security and access management solutions for data assist data-centric organizations in minimizing their operations, facilitating increased protection, and at the same time allowing the utilization of their value contained in the data without experiencing delays. In addition, the use of dynamic data security may reduce policy management to the complexity of obtaining appropriate data to the right persons in due time. This makes it much easier and more efficient for data collaboration across business units, geographic locations, and with external parties. This includes the ability to rapidly develop and deploy data products that impact the bottom line in a positive way [40].
Continuous Monitoring

Immuta delivers quick awareness to risky user behaviour regarding user data access and that facilitates data security posture management. Continuous monitoring and detection capabilities enables you to continuously track the usage of sensitive data and identify the user involved and at the same time uncover the vulnerabilities withing your data environment [41].

[4] `https://www.immuta.com/product/`

Immuta's data masking

Immuta's dynamic data masking offers a solution for organizations in various sectors to streamline their processes, enhance data protection, and realize the value of sensitive data. Through the application of dynamic data masking, you have the ability to distribute data in adherence to laws, standards, and agreements on data sharing, lessen the technical workload through dynamic policy enforcement without the need to duplicate data [42].

Advantages of using Immuta data masking can be summarized as:

- **Plain Language Policy Authoring**: Immuta enables the creation of attribute-based access control policies either in straightforward language or through coding, making them manageable for stakeholders of any level. This facilitates data masking without requiring specialized technical expertise.

- **Dynamic data masking**: Manually duplicating data and then redacting or anonymizing it can slow down analysis and diminish the usefulness of the data. Immuta's dynamic data masking policies offer techniques such as hashing, regular expression, rounding, conditional masking, and substitution with null values or constants, including options for reversible and format-preserving masking, as well as external masking. This is achieved without the need to copy or relocate data.

- **Condition data masking**: Guard against data breaches without depending on manual modifications. Immuta automates the implementation of access limits using masking policies that are contingent on specific conditions, including time periods, the geographical locations of users, and data found in neighboring cells or linked tables. Through the use of conditional logic in its masking policies, Immuta provides adaptable policy application while minimizing exposure to risk.

4.2.5 Apache Ranger

Apache Ranger[5] is a free, open-source security framework designed to enhance security on big data platforms, including Hadoop. It acts as a security guard for your data, managing and enforcing comprehensive access controls and security policies across the data ecosystem. Apache Ranger offers a centralized system for managing security policies, including user and group access controls, authorization protocols, and audit regulations. Its integration with prevailing big data platforms allows for the consistent application of these security measures throughout the entire big data infrastructure [43].

How Ranger works

Apache Ranger operates through a central administration interface, allowing the creation, modification, and implementation of security policies. These policies can be tailored based on several factors, such as the identity of the user, IP address, the time of access, and the sensitivity of the data involved [44].

[5] `https://ranger.apache.org/`

The most important Apache Ranger use cases

- **Role-based Access Control Enforcement(RBAC):** Apache Ranger facilitates the development and oversight of RBAC policies tailored for big data environments.
- **Data Masking:** Apache Ranger is capable of obscuring confidential information according to the user's level of authorization, thereby preventing access by unapproved personnel.
- **Data Encryption:** Apache Ranger offers policies for encrypting data, safeguarding it while stored and during transfer.
- **Activity Monitoring and Audit Trails:** Apache Ranger delivers comprehensive audit records, detailing user interactions and access to sensitive information.

Row Level filtering and Column masking

Apache Ranger policy model has been enhanced to support row-filtering and data-masking features.
Column Masking Overview
Column-masking provides a secure and adaptable way to quickly conceal sensitive data in a Hive source. Through security policies managed by Apache Ranger, you can dynamically mask or obscure personal information at the column level within Hive query results. By employing various masking techniques, it's possible to configure a column to show only the year from a date, or just the first or last four digits of a number, among other options. Column-masking is governed by several rules:

- Masking can be assigned to individual users, specific groups, and under certain conditions.
- A unique masking policy is required for each column.
- The sequence in which masks are applied follows their arrangement in a query or within a security policy.

Ranger-based column masking operate as a form of "implicit view," substituting references to tables or views in an SQL query before the query is executed. This substitution is determined by analyzing user permissions. For instance, if a user has access to "table1" but is subject to a masking rule on "table1.column" that changes its content to 'xxx'. An example of this feature could be that this:

```
SELECT column_1, column_2, column_3
FROM table1
```

would become this:

```
SELECT 'xxx' AS column_1, column_2, column_3
FROM table1
```

Row level filtering overview

Row-level filtering simplifies queries and bolsters data security for user- or role-based query executions By employing SQL functions or Apache Ranger security policies, access can be confined at the dataset level, influencing the execution of queries. Implementing row-level security on compatible tables serves to minimize the risk of sensitive data exposure to certain users or groups, thereby enhancing control over data access.

Row-level restrictions may be set by user, group/role, and other conditions. This is a topic that will be very well studied in the other chapters of the thesis as it holds a very important role in the project developed.

An example would be:

```
SELECT column_1
FROM table_1
WHERE column_3
```

This one would become:

```
WITH filtered_table_1
AS (
  SELECT column_1, column_2, column_3
  FROM table_1
  WHERE column_2
)
SELECT column_1
FROM filtered_table_1
WHERE column_3;
```

In this example above, the table has been filtered in the first place with the column_2 before the second query of ours have the occasion to query it.

CHAPTER 5

PROJECT IMPLEMENTATION

As this Thesis derive from work done within a collaboration with a company; we had to create a deliverable requested by the client. In this chapter, I am going to work you through the project implementation from the beginning to the end. I will explain what was the client's request in more details and how we executed it.

5.1 The client's requirements

The client's objective was to implement a type of data protection through dynamic data masking with specific operating logic.

The user observing the database had to access in 2 different modalities; RBAC and ABAC.

5.1.1 Role Based Access Control

Role-based access control (RBAC) is a system that manages access based on the roles assigned to users. It grants permissions and access rights to users based on their roles within an organization.

In most organizations, access to sensitive data is allocated according to the varying roles and responsibilities of employees through RBAC. This method ensures that only staff members holding specific roles are authorized to view or modify certain pieces of information.

Access rights for each employee are defined by their respective roles, determining the level of permissions they have within the system. For instance, access can be tailored so that only managers, testers, or basic users can interact with particular resources or perform certain operations, depending on their role [45].

5.1.2 Attribute Based Access Control (ABAC)

Unlike RBAC, which grants access based on user roles, ABAC (Attribute-Based Access Control) relies on the user's attributes to determine their access rights to resources. The characteristics of the user requesting access are critical in making this decision.

ABAC provides a more granular level of access control, focusing on detailed attributes rather than broad roles, thus offering a deeper and more nuanced approach to managing access [46].

How does ABAC work

In order to allow access , ABAC checks the requester's attributes. These attributes can be subdivided into four main parts:

- **Subject/user**: attributes defines the user making an access request. Examples include name, codice fiscale, job title, the user's role, organization the requester work for, department he is assigned to and his security clearance.
- **Resources/object**: attributes are the description of the resources being accessed.
- **Action**: is the type of activity the user will do to the resource requested. Examples include Read, Write, Update, delete, etc.
- **Environmental**: attributes describe the conditions when the user attempts to access the resource. Examples include time and location, device used, etc.

An ABAC system checks the environment attributes and rules are developed that will clarify what attributes to be met in order to allow access a user [47].

In summary, an ABAC system creates policies and rules to design a combination of different attributes(subject, environmental,user). These attributes are needed to perform an action to a resource. Grant and deny accesses are provided by those policies created.

For example, if you do not want the bank employees to view all the data, ABAC can place limitations so only employees of the bank in Turin can view sensitive information.

Here is how it works:

When an access request is generated, attributes are checked by the ABAC tool to see if they match established policies. If they match, the user will have access to a resource. In our case, it would be bank employees who works Turin, and who is trying to view sensitive information.

Figure 5.1 shows in details how in our project we implemented the ABAC system for better granularity and a high level of control.

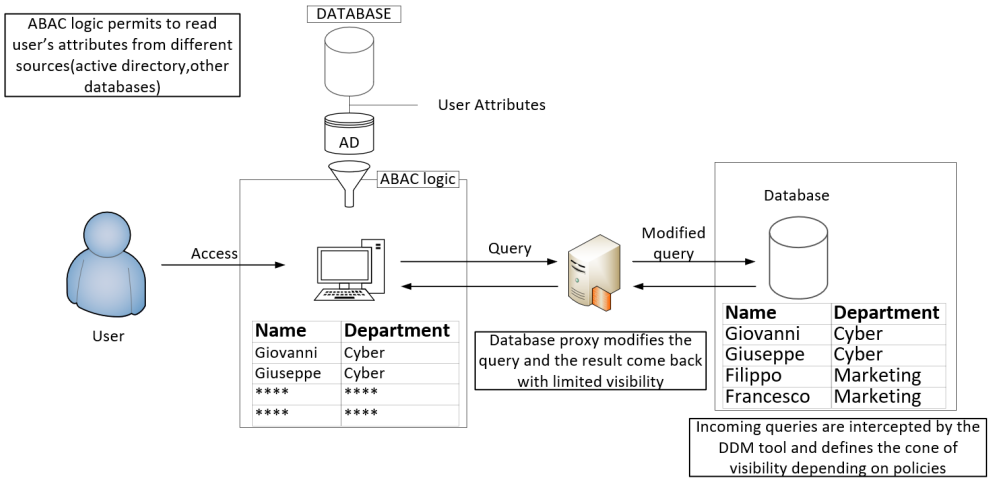


Figure 5.1: Project Architecture

5.2 User Attributes

In the course of our project, ensuring and enforcing security protocols required pinpointing the locations of user attributes/information. The client relied on two distinct data sources to gather this information.

5.2.1 Active directory

it is a logical grouping of users and assets within a domain. For each user, his detailed information is recorded, including:

- User Memberships
- Name, Surname
- Email
- Other context information (account expiration date, account creation date, etc.)

5.2.2 Secure database

This database is the other data source which contains part of the user's attributes as well as the mapping between the acronym and related enabling groups.

In this database we get all the necessary information used by policies implementing rules as a gauge.

Some of the information retrieved are:

1. **User role**: This is the type of user trying to connect.
2. **Legal Entity**: This is self explanatory as it is the legal entity the user is affiliated to.

Both of these information are highly important in the implementation of the legal entity policy. This will be further explained in later chapters.

5.3 Policies

Based on the user's attributes, the Dynamic Data Masking solution authorizes the user to view the data, otherwise it applies the masking policies.
We are going to take a look at those masking policies in details as they are the pivotal point we base this whole project on.
Policies are access and visibility rules that are implemented by a tool on the data. They are subdivided into 5 categories and below we are going to explain each category in detail.

Those categories are:

- ACL POLICY
- GDPR POLICY

- LEGAL ENTITY POLICY
- RESERVED POLICY
- PROFILED POLICY

5.3.1 ACL POLICY

The objective of the Access Control list policy(ACL) is to provide access to data based on the user's attributes, in this case guaranteeing access only if in possession of the ACL enabling code associated with the queried acronym(database). This policy inhibits users who access any database they don't have permissions to.

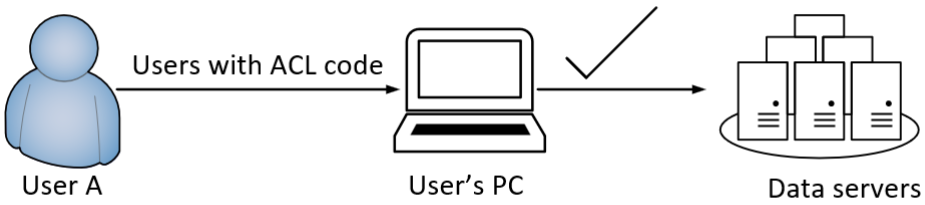


Figure 5.2: User with ACL codes

Figure 5.2 shows that when users have an ACL code for the database, they can access them.

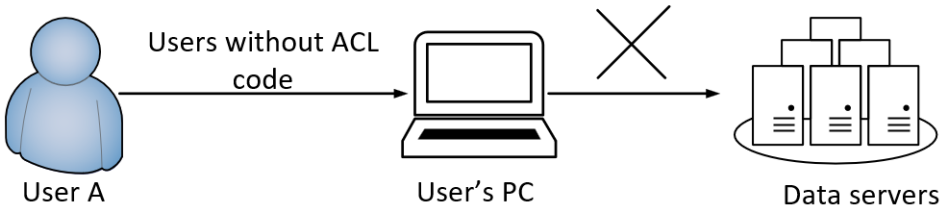


Figure 5.3: User without ACL codes

Figure 5.3 on the contrary shows how a user with no ACL code has no possibility to access data in hive servers.

In the implementation of the ACL policy, each acronym has to be defined with its appropriate ACL groups. Users found in those groups are the ones allowed to access that acronym. By default, the masking tool blocks all users not belonging to the indicated groups who try to access the resource.

ACL FLOW

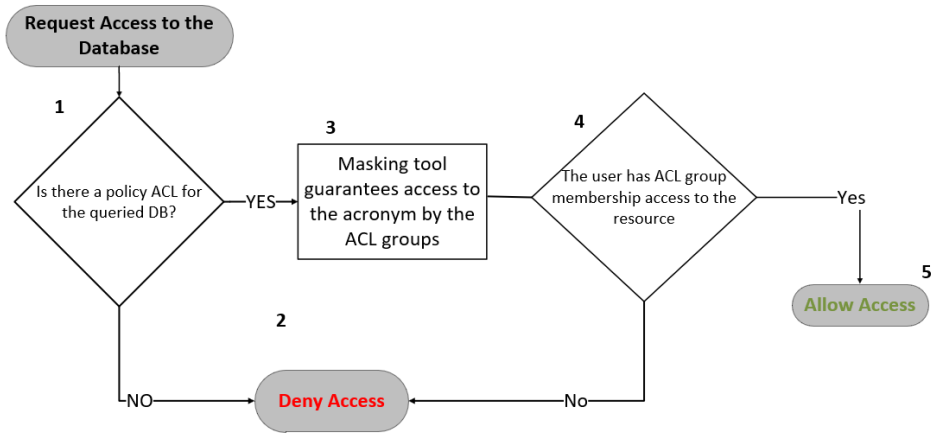


Figure 5.4: Flow of the ACL implementation

In Figure 5.4 is presented the ACL flow implemented, which consists in the following actions:

1. The masking tool verifies if it exists an ACL policy associated to the acronym (DB) that the user is requesting.

2. The masking tool by default applies a Deny access if there is no ACL policy associated to the requested resource.

3. The tool finds the ACL policy and guarantees that the access is through ACL groups.

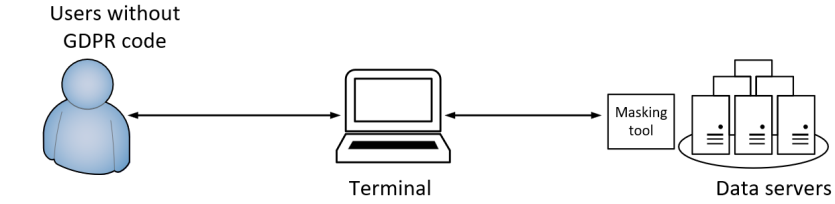
4. The tool checks if the user connected belongs to at least one of those ACL groups. (These groups are part of the user Attributes and can be found on the Access Directory.)

5. The tool gives access of that acronym(DB) to that user if the previous check was a success otherwise the user is denied.

5.3.2 GDPR POLICY

The objective of the GDPR policy is to provide protection of data intended as personal, Indicated by the client, based on the user's attributes. In this case, guarantee the visibility of personal data only to those in possession of the GDPR enabling code associated with the acronym in question.

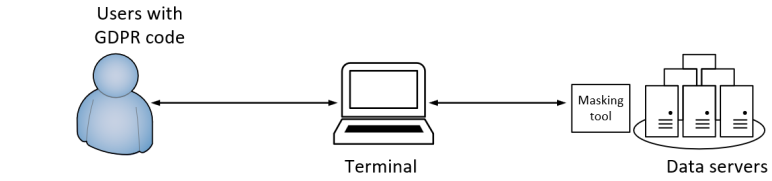
Figures 5.5 and 5.6 shows the 2 different users one with gdpr code who accesses the database in clear and the other one without the code who only sees sensitive data masked.



ID	Name	Email
1	XXXX	XXXX
2	XXXX	XXXX
3	XXXX	XXXX
4	XXXX	XXXX

ID	Name	Email
1	Mario	Mario@tirocinio.com
2	Maria	Maria@tirocinio.com
3	Chiara	Chiara@tirocinio.com
4	Anna	Anna@tirocinio.com

Figure 5.5: unauthorized user to view the data



ID	Name	Email
1	Mario	Mario@tirocinio.com
2	Maria	Maria@tirocinio.com
3	Chiara	Chiara@tirocinio.com
4	Anna	Anna@tirocinio.com

ID	Name	Email
1	Mario	Mario@tirocinio.com
2	Maria	Maria@tirocinio.com
3	Chiara	Chiara@tirocinio.com
4	Anna	Anna@tirocinio.com

Figure 5.6: authorized user to view the data

GDPR FLOW

Ranger, our masking tool allows you to develop two types of policies: resource based and tag based. Resource Based Policies are applied to resources that are specified in the policy itself (DB/table/column).

Tag Based Policies are applied using, as the name suggests, a tag. The tag allows you to control multiple resources with a single policy.

Ranger allows you to implement a single masking policy for each acronym, but having the need to implement masking on multiple columns of multiple tables within the same acronym, this could not be done with a normal resource based policy.

To deal with this problem, the GDPR policy is developed following the tag based policy method and is defined as follows:

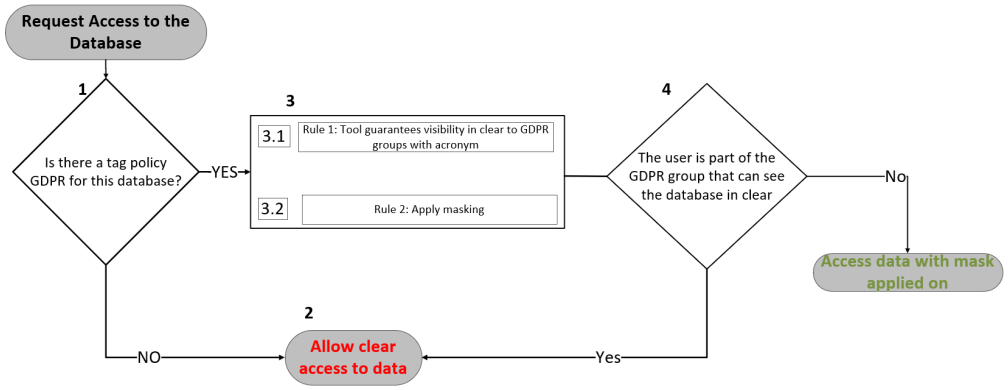


Figure 5.7: Implemented GDPR Flow

In Figure 5.7 is presented the GPDR flow implemented, which is composed of the following actions:

1. The masking tool checks if there are any policies (Tag based) associated with the table that the user is querying.
2. The tool does not find policy tags associated with the queried tables and allows clear access on all columns.
3. The masking tool finds a masking policy associated with the requested table and runs the rules:
   (a) Grants clear visibility to users who belong to the following group: GDPR.
   (b) Mask all other users.
4. The tool checks if the user belongs to the GDPR group, if not, the rule described in 3.b is applied, while, if so, the rule 3.a is applied.

With GDPR, we have labels to protect, each of which identifies a type of data. The security product will have to protect these types of data with a masking logic. Some of the labels are as follows:

1. GDPR_DATO PARTICOLARE: This includes
   - BIOMETRICAL DATA: e.g., image shape, iris image...
2. GDPR_DATO ANAGRAFICO: This includes
   - NAME AND SURNAME e.g., John Smith,...
   - EMAIL e.g., mario@tirocinio.com
   - CODICE FISCALE e.g., QWEDFV76L45W480B
   - HOME ADDRESS e.g., via Verdi, 5
   - DATE OF BIRTH e.g., 05/03/1934
3. GDPR_DATO BANCARIO E OPERATIVO: This include
   - IP ADDRESS e.g., 138.147.34.217

TYPE OF MASKING

The type of masking done a value depends on what is the type of that data before being masked. We have two different ways of masking.

- **String masking**: string masking is used every time we try to mask a string value. At that point the mask is **XXXX**.
- **Date masking**: date masking is used every time we try to mask a date value. At that point a data changes from **Day/Month/Year** to ****/**/YEAR** in which YEAR is the original value of the date.

5.3.3 LEGAL ENTITY POLICY

In the context of the definition of access control policies to the data present in the acronyms. The policies for the Legal Entities aim to define and implement the cones of visibility, limiting the visible information only to the data users based on the legal entity to which they belong and the user role they possess, in line with the principle of least privilege.

We have 2 scenarios in our legal entity policy, the first one is of a user with role A who has only visibility to data of the legal entity he belongs. The other scenario is a user with role B who has an entire visibility of the data independently of the legal entity.

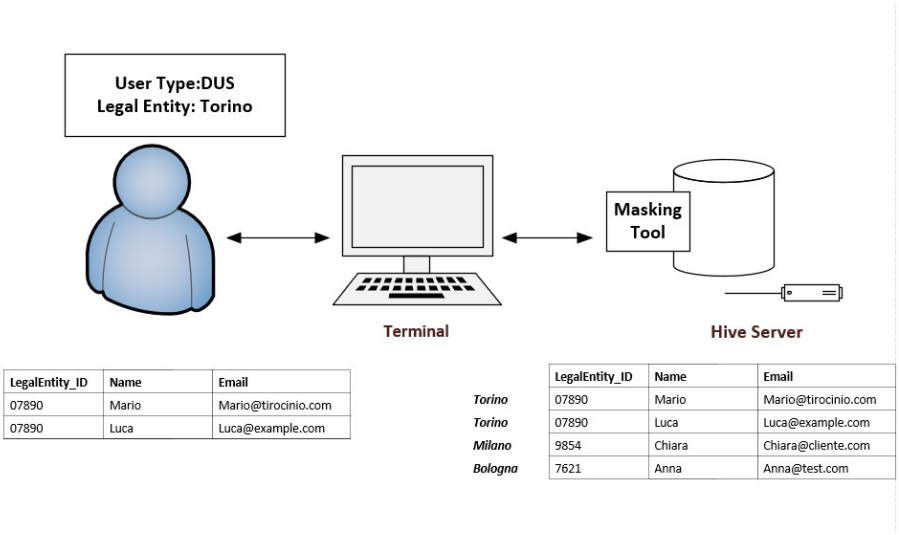


Figure 5.8: user with Role A and visibility only on Torino Legal entity

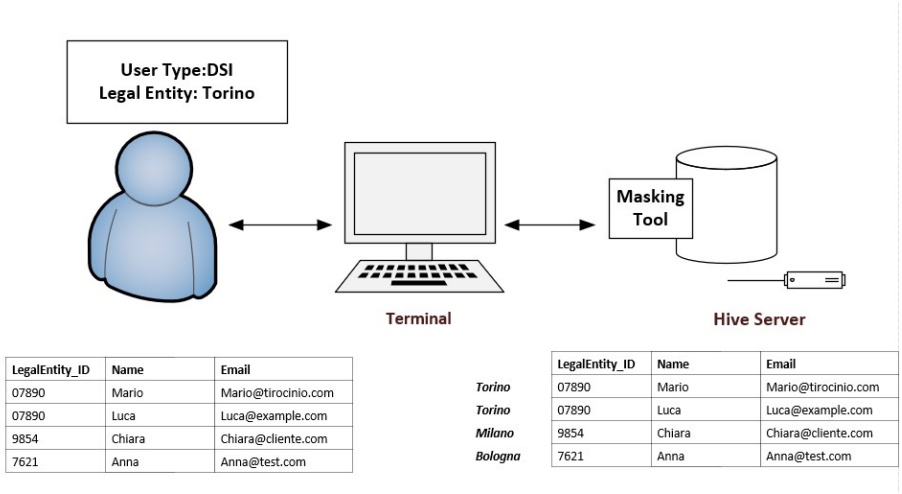


Figure 5.9: User with Role B and authorized to view all the data

Figures 5.8 and 5.9 above show how a user when connecting to hive servers receives the information. Other users of a different legal entity will see their data accordingly.

LEGAL ENTITY FLOW

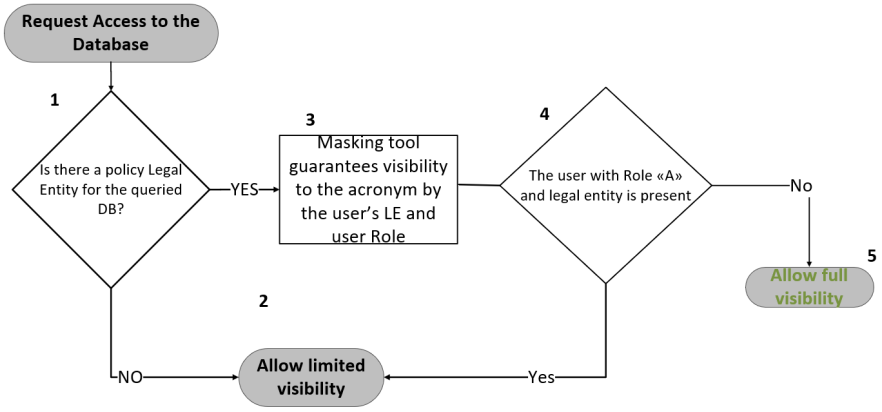


Figure 5.10: Legal Entity flow graph

In Figure 5.10 is present the Legal Entity flow implemented, which consists of the following actions:

1. The masking tool verifies if it exists a LE policy associated to the acronym (DB) that the user is requesting.
2. The masking tool by default applies a least visibility access if there is no LE policy associated to the requested resource.
3. The tool finds the LE policy and guarantees that the access is through user Role and Legal Entity.
4. The tool checks if the user connected has User Role "A" and Legal Entity (the User Role and Legal entity are user Attributes and can be found on the secure database).

5. The tool gives partial visibility of acronym(DB) to that user if the previous check was a success otherwise the user has full visibility.

5.3.4 RESERVED POLICY

The objective of the Reserved policy is to provide protection of data intended as Confidential, Indicated by the client, based on the user's attributes. In this case, we guarantee the visibility of personal data only to those in possession of the Private Data enabling code associated with the acronym queried.

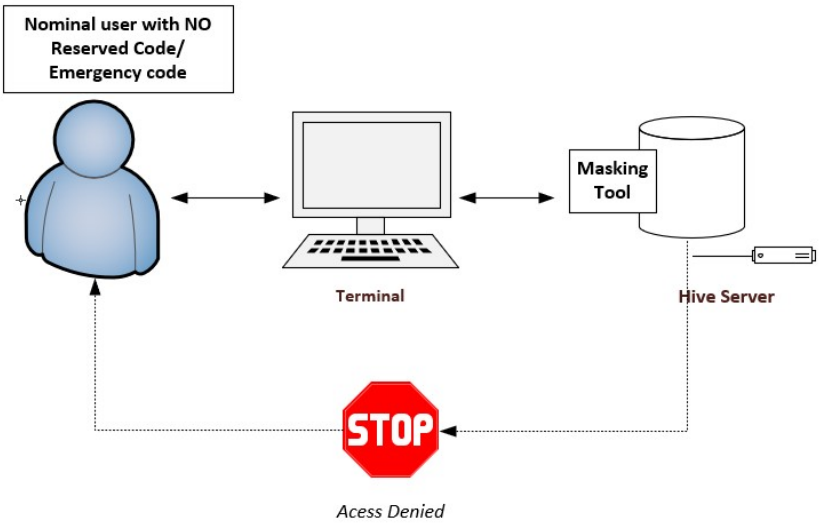


Figure 5.11: Reserved Policy with a user without a Code

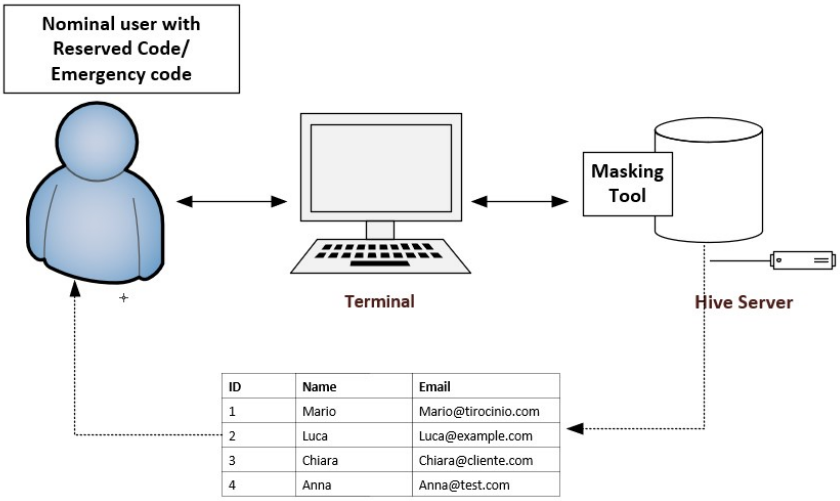


Figure 5.12: Reserved Policy with a user with a code

Figure 5.11 and 5.12 presents the reserved policy in different case where a user does not have abilitating code on the first image hence could not access the table,

on the contrary, the latter image shows a user with the code that allows him full access to the database.

The reserved policy is like the acl policies but instead done on the table instead of the whole database.

Reserved Data flow Graph

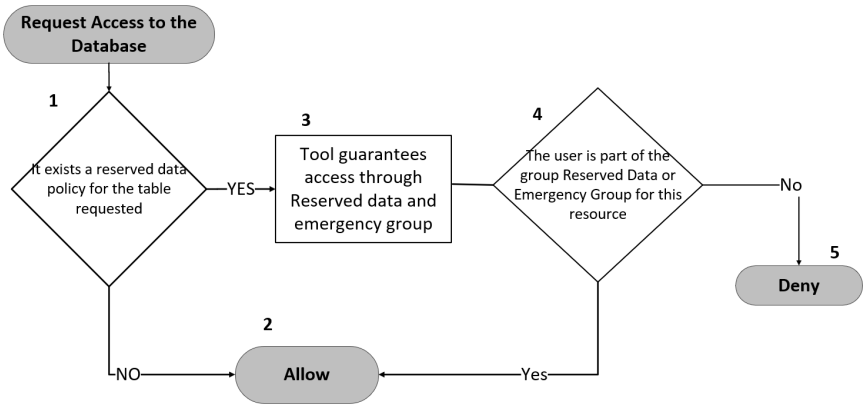


Figure 5.13: Reserved Policy control flow

In Figure 5.13 is presented the implemented reserverd policy control flow, which is composed of the following actions:

1. Ranger(Our dynamic data masking) checks if there are any policies (tag based) associated with the DB that the user is querying;
2. Ranger allows access if there is no Confidential Data policy associated with the requested resource;
3. Ranger finds an access policy associated with the requested resource and grants access to Confidential Data (DR) or Emergency groups
4. Ranger checks if the user belongs to one of the groups associated with the resource access policy

5.3.5 PROFILED POLICY

The Profiled Data policy is a Custom policy that reflects the client's need to protect the data of its acronyms based on the values contained in the tables and on the memberships associated with the user's profile.

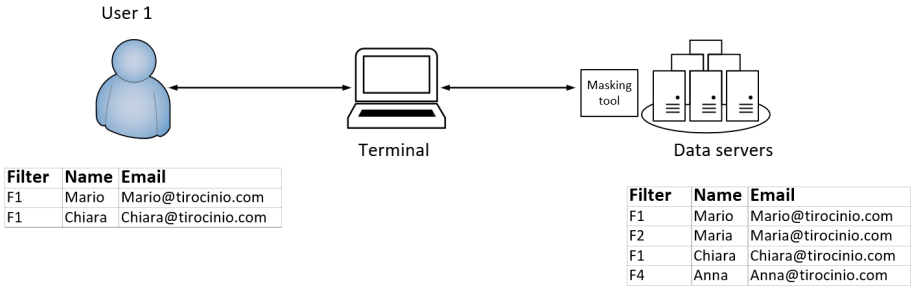


Figure 5.14: Profiled Policy

Figure 5.14 presents a filtering case where it depends on the value in the table, in our example we filtered depending on the user having a value F1 in a column.

The Profiled Data policy is divided into two different types of policies:

- Consent Policy
- Policy legal nature

Consent Policy work flow

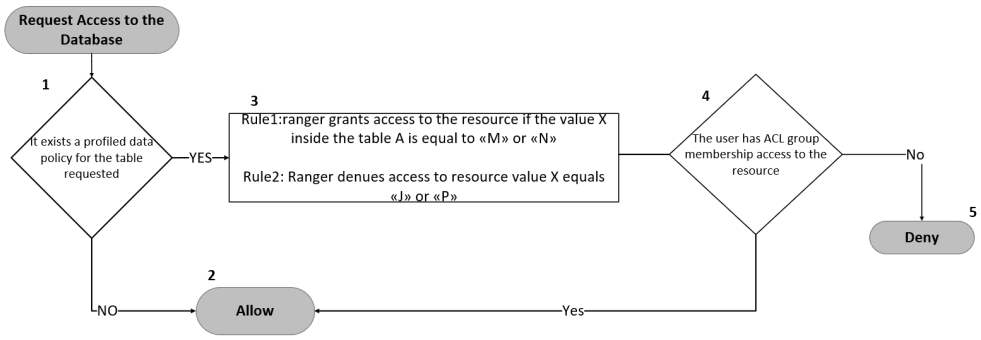


Figure 5.15: Implemented Consent Profiled Policy

Figure 5.15 presents the implemented consent profiled policy, which consists in the following actions:

1. The data masking tool (Ranger) checks if there are any policies (resource based) associated with the DB that the user is querying;
2. The tool allows access if there is no Profiled Data policy associated with the requested resource;
3. Ranger finds an access policy associated with the requested resource and checks if two filtering rules
   (a) If the value “X” value contained in the “A” table is equal to “M” or “N” it grants access to the resource
   (b) If the value “X” value contained in the “A” table is equal to “J” or “P” it denies access to the resource

4. If the value “X” value contained in the “A” table is equal to “M” or “N” it allows access to the resource, otherwise it denies it

Policy legal nature work flow

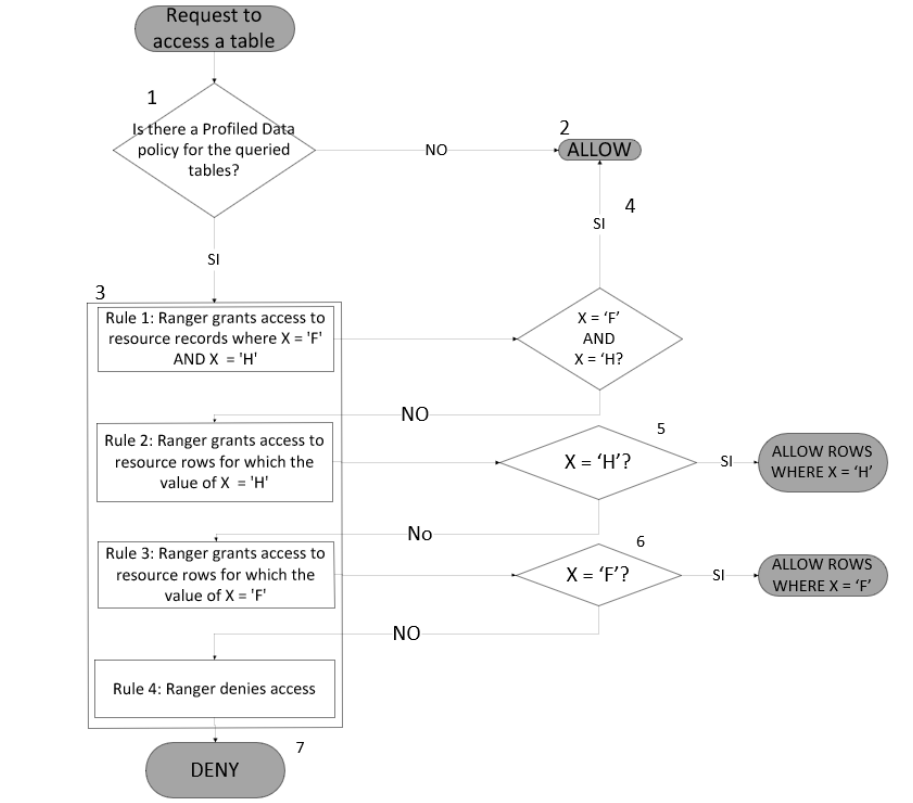


Figure 5.16: Legal Nature Policy Workflow

The legal nature policy on Ranger is developed following the Resource Based Policy method. Figure 5.16 presents the implemented policy, which is defined as follows:

1. The user queries our database;
2. Ranger allows access if there is no Profiled Data policy associated with the requested resource;
3. Ranger finds a Profiled Data policy associated with the requested resource and checks which rule to apply:
   (a) The user belongs to both group H and group F
   (b) The user belongs to Group F
   (c) The user belongs to Group H
   (d) None of the above rules apply
4. If the user belongs to both group F and group H, he has access to the resource
5. If the user belongs to Group F, he has access to the lines where X = ’H’

6. If the user belongs to Group H, he has access to the lines where X = 'F'

7. If none of the previous rules are applied, access is denied and ranger checks if there are any policies (resource based) associated with the DB

5.4 How Were the Policies Implemented

The question anyone would ask himself at this point would be, how is it possible to keep up with the policies creation and not be overwhelmed or confused when creating them?
The simple answer to that question is that it would be very difficult to create the policies manually.
We developed a tool that were connecting to ranger APIs and through a series of configuration files and properties file, create those policies for each acronyms/database.
There was also a need to create another plugin called the Hook along the development of the Legal entity policies.
This was due to the fact that our legal entity was to be customized to take the values of the user to be filtered from the database.

CHAPTER 6

THESIS CONCLUSION AND FUTURE WORK

This thesis has explored the critical realm of data protection in an era where digital information's value and vulnerability are both at their peak. Through a comprehensive examination of data breaches, the impact they wield, and the subsequent legal frameworks developed to mitigate these risks, we've delved into the intricacies of safeguarding sensitive data.

We also saw how dynamic data masking (DDM) plays a pivotal role for fortifying data privacy within the financial services sector, particularly in light of the European Union's stringent data protection mandate, GDPR. Through a comprehensive investigation and a practical project implementation at PwC Italy, this study has illuminated the essential role of DDM in transforming how businesses manage and secure user data.

Our results show that dynamic data masking (DDM) works well in meeting the GDPR's rules for limiting data use and helps create new, higher standards for protecting privacy across different industries. Despite facing challenges such as integration complexities and the need for alignment with evolving compliance requirements, the project highlighted the indispensable value of combining technical solutions with organizational policies to achieve robust data protection.

As of our project right now, we are evolving with analysing how DDM can work in the new European union regulation on financial service called DORA(Digital Operational Resiliency Act).

A promising direction for future research lies in the evolution of data masking tools through the integration of artificial intelligence (AI). Currently, the implementation of dynamic data masking (DDM) requires the meticulous creation of policies for each data column to be masked, a process that can be both time-consuming and complex, especially in large databases with diverse data types. An AI-enhanced tool could revolutionize this process by automatically analyzing the data within each column, identifying sensitive information such as personal identifiers (e.g., Social Security Numbers, credit card details, and so on) based on patterns, context, and predefined criteria. By recognizing the nature of the data, such a tool could then apply the most appropriate masking technique without the need for extensive manual policy configuration. This advancement would not only streamline the setup and maintenance of DDM systems but also significantly enhance the efficiency and adaptability of data protection measures in response to evolving regulatory require-

ments and data landscapes.

In conclusion, this thesis highlights how important dynamic data masking is for protecting data today. It provides a detailed way to keep sensitive information safe from unwanted access and breaches.

BIBLIOGRAPHY

[1] ThriveDX, https://thrivedx.com/resources/article/data-breach-types

[2] IBM, https://www.ibm.com/downloads/cas/3R8N1DZJ

[3] flashpoint, https://flashpoint.io/blog/what-are-data-breaches-how-to-prevent/

[4] upguard, https://www.upguard.com/blog/cost-of-data-breach

[5] Biggest data breach in the financial industry, https://www.upguard.com/blog/biggest-data-breaches-financial-services

[6] cybersecuritydive, https://www.cybersecuritydive.com/news/first-american-financial-offline-cyber-incident/703262/s

[7] SEC, https://www.sec.gov/news/press-release/2021-102

[8] CalgaryHerald, https://calgaryherald.com/news/local-news/class-action-lawsuit-claims-city-leaked-personal-information-of-3700-employees.

[9] DW, https://www.dw.com/en/deep-root-analytics-behind-data-breach-on-198-million-us-voters-security-firm/a-39318788

[10] SolarWind, https://www.techtarget.com/whatis/feature/SolarWinds-hack-explained-Everything-you-need-to-know

[11] SolarWindcost, https://www.cybersecuritydive.com/news/solarwinds-1-year-later-cyber-attack-orion/610990/#:~:text=For%20SolarWinds%2C%20the%20newly%20minted,quarterly%20report%20from%20October%20said.

[12] J. Wolff and N. Atallah, "Early gdpr penalties: Analysis of implementation and fines through may 2020", Journal of Information Policy, vol. 11, 2021, pp. 63–103

[13] Statista, https://www-statista-com.ezproxy.biblio.polito.it/statistics/1172445/countries-with-highest-fines-issued-gdpr/

[14] DataPrivacyManager, https://dataprivacymanager.net/5-biggest-gdpr-fines-so-far-2020/

[15] C. Goyal, "Data masking: need, techniques & solutions", Int. Res. J. Manag. Sci. Technol.(IRJMST), vol. 6, no. 5, 2015, pp. 221–229

[16] imperva, https://www.imperva.com/learn/data-security/static-dynamic-data-masking/

[17] "what is data masking needs,techniques & best practices." https://blogs.bmc.com/data-masking/?print-posts=pdf

[18] techtarget, https://www.techtarget.com/searchsecurity/definition/data-masking

[19] https://www.dataversity.net/data-masking-best-practices-and-benefits/

[20] R. Archana and R. S. Hegadi, "Applications of data masking techniques for data security", 2014

[21] H. Li, L. Yu, and W. He, "The impact of gdpr on global technology development", 2019

[22] https://gdpr.eu/what-is-gdpr/

[23] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[24] https://www.cdc.gov/phlp/publications/topic/hipaa.html

[25] https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

[26] https://legal.thomsonreuters.com/en/insights/articles/understanding-california-consumer-privacy-act#:~:text=The%20CCPA%20grants%20consumers%20several,personal%20information%20the%20business%20holds.

[27] https://www.acc.com/resource-library/quick-overview-understanding-california-consumer-privacy-act-ccpa

[28] https://oag.ca.gov/privacy/ccpa#heading4

[29] https://www.onetrust.com/blog/the-ultimate-guide-to-pipeda-compliance/

[30] https://laws-lois.justice.gc.ca/eng/acts/p-8.6/page-7.html

[31] https://www.bath.ac.uk/legal-information/data-protection-act/

[32] https://www.imperva.com/learn/data-security/gdpr-article-32/

[33] https://www.linkedin.com/advice/3/you-need-protect-sensitive-data-analysis-how-rgpfc

[34] https://maskingdocs.delphix.com/

[35] https://geekflare.com/data-masking-tools/

[36] https://docplayer.net/19482698-Data-masking-a-white-paper-by-k2view-abstract-k2view-data-masking.html

[37] https://www.informatica.com/content/dam/informatica-com/en/collateral/data-sheet/dynamic-data-masking_data-sheet_1779.pdf

[38] https://www.immuta.com/product/

[39] https://www.immuta.com/product/discover/

[40] https://www.immuta.com/product/secure/

[41] https://www.immuta.com/product/detect/

[42] https://www.immuta.com/product/secure/dynamic-data-masking/

[43] https://www.dremio.com/wiki/apache-ranger/

[44] https://docs.dremio.com/current/sonar/security/rbac/row-column-policies-ranger/

[45] https://www.imperva.com/learn/data-security/role-based-access-control-rbac/

[46] https://www.sailpoint.com/identity-library/what-is-attribute-based-access-control/

[47] https://blog.satoricyber.com/an-introduction-to-attribute-based-access-control/