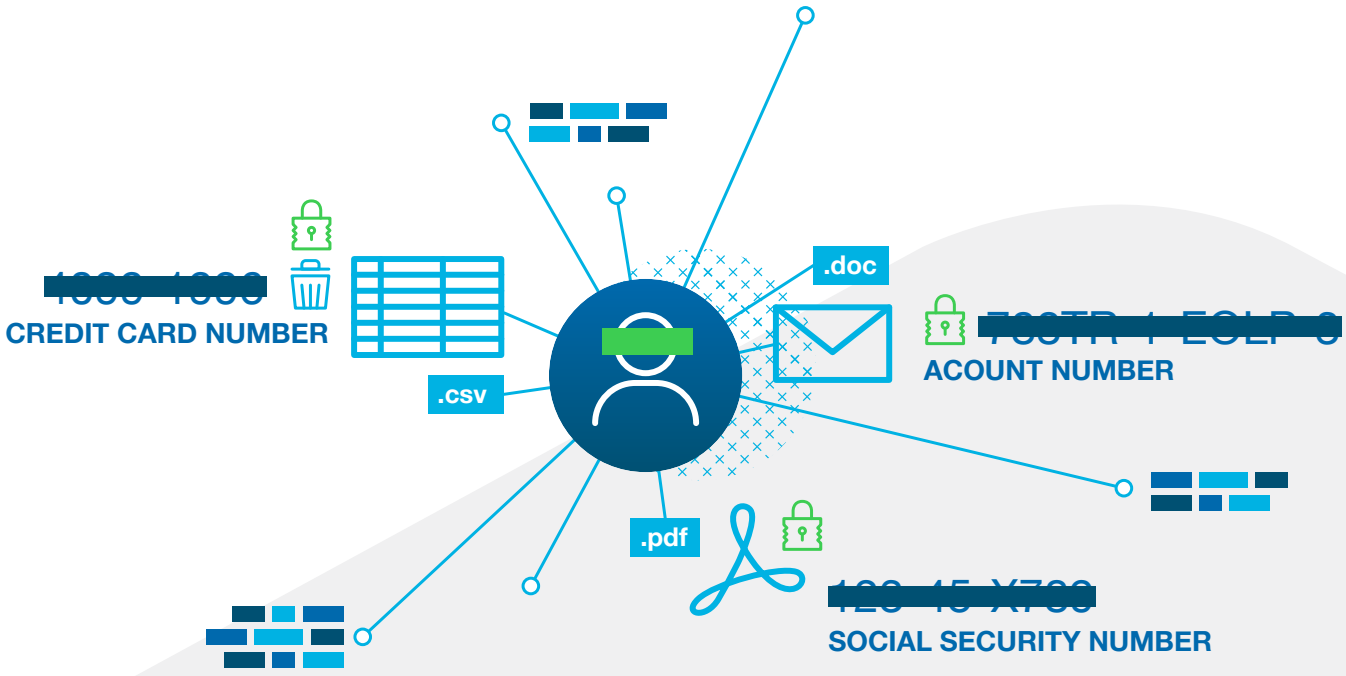
SPIRION

WHITE PAPER

Data De-identification

Privacy challenges and Business Utility



Introduction

One of your company's greatest assets lies on your servers and storage devices – your data. Each piece of information you have collected from your customers, employees, vendors, and devices tells the story of your business's past. By strategically using the collected data, your business can make accurate decisions that work to grow the company. However, Forrester found that 60 to 73 percent of all enterprise data goes unused for analytics.

Companies must gain consumer trust to gain access to the data needed to run their business, which is often an uphill battle. According to Pew Research, 81% of consumers say that the risks of data collection by companies outweigh the benefits. When companies are breached or fined for data violations, consumers are less likely to give access to data, making it harder to gain the insights needed to grow the business.

With stringent privacy regulations, such as CCPA, HIPAA, and GDPR, a business must comply or risk hefty fines and permeant damage to the brand's reputation. Each time your business uncovers new sensitive data during a discovery scan, you must classify the data and then take a remediation action to comply with privacy laws. Additionally, while you may opt to protect the data, the legislation allows for consumers to request their PII be permanently deleted from all the company's data storage locations.

Businesses have no choice but to comply or risk costly fines – as well as the damage to their reputation. Additionally, breaches require significant time and effort to resolve – with IBM reporting the average breach took 73 days to contain. Some companies proactively reduce their risk for breaches by taking permanent remediation action even without a customer's request.

Remediation Without Losing Business Value

When faced with the remediation decision – either by request or proactively – many companies automatically take a permanent approach to the data's removal. Common actions, such as shredding or deleting, result in the company losing all the data.

Other companies turn to redacting data, which means altering it to the point that it is non-identifiable, yet still usable. Another common option is encryption, which was designed to protect data during transmission. However, this route does not allow the business to use the data because it cannot be accessed without the encryption key, which violates data privacy regulations.

While the only truly safe data is deleted data, we need data for further innovation and research. For example, do we want scientists studying a pandemic without data? Every time you delete data, your business can no longer compare data on a year-by-year basis, because you don't have access to data from customers who are not currently active. Businesses lose the ability to learn from past successes and mistakes without this ability to compare critical metrics.

A cornerstone for business is retaining customers and understanding customer churn. However, companies must have lapsed or inactive customers in their database to compare the information with current customers. When businesses take removal action on data, such as shredding, redacting, deleting, or encrypting, they are unable to make often simple changes that could increase customer retention and company revenue.

Over time, losing customer data severely hampers an organization's ability to understand customers, grow the business, and maintain productivity. The harm to the business increases further when competitors have access to critical data that allows them to proactively predict customers' needs for marketing, sales, and product production. As customers' expectations increasingly rise with many companies having extensive data, companies without data fall further behind.

Using De-identification to Comply with Privacy Regulations and Retain Business Value

Companies of all sizes and in all industries are increasingly turning to de-identification for remediating sensitive data. According to an Egress survey, 93% of US IT decision-makers said they had at least taken some steps to comply with privacy regulations such as CCPA or the GDPR. De-identification is a form of differential privacy that removes all identifying features but retains enough information to hold business value. The resulting data satisfies the privacy legislation requirements, so you can legally retain the data in the altered form for business purposes.

To de-identify data, technology data scientists evaluate and then modify the three critical components of data – direct identifiers, indirect identifiers, and safeguards/controls. A direct identifier is a piece of data that identifies the person without additional information, such as name or social security number. Indirect identifiers help connect other pieces of data together, such as date of birth and gender, to identify the person.

In the de-identification process, the direct and indirect identifiers are either removed or modified to the point that the person's identity cannot be discovered. For example, customer data may be de-identified by removing names, SSNs, and credit card information, as well as changing random genders and modifying birthdates.

De-identification cures the data in a way that the business retains the majority of the utility of the data while meeting data privacy regulations, such as GDPR, HIPAA, and CCPA. By using de-identification technology, companies can ensure their data meets all legislation required for the data type and location of the company.

De-identification Balances Privacy with Utility

For many organizations, de-identification is the perfect balance of privacy and utility. Businesses often see the following benefits from de-identification:

- **Variety of uses for the data** - Unlike other forms of remediated data, such as encryption, de-identification satisfies many business requirements instead of a single-use case, including mitigating privacy risks, data sharing, and data security.
- **Retains the business value of the data** - The business retains the use of the data, which can be immensely valuable when used correctly. Additionally, organizations do not lose their competitive edge to other businesses that retain the business value of their data.
- **Use the data for other purposes** - After the de-identification process is complete, the business can use the data for gaining insight, testing, resolving errors, machine learning, and marketing. The data can also still be used to accurately show business growth and customer churn.
- **Meets privacy grade** - Because de-identification ensures that the data includes privacy management, privacy control, and security capabilities, the de-identified data meets privacy grade requirements. And customers are more likely to trust companies with their data and business that consistently ensure privacy grade protection.

Using De-identified Data to Gain Business Value

Many businesses fail to take advantage of their valuable data for growing their business. But companies that proactively look for use cases and applications for their data can harness that data – which is a unique asset their competitors do not have.

A common use of de-identified data is creating testing environments for analyzing large data sets. Privacy laws restrict companies from performing tasks with data containing credit cards. However, once the data undergoes de-identification, the data meets the privacy regulations and still provides enough value to conduct tests for new applications.

By using data for machine learning, companies increase the value of their data as they can create smarter applications and even more data. This allows organizations to quickly compound and grow the value of their data even further. Through machine learning, companies can more accurately predict customers' needs, preferences, and next actions, which allows them to proactively meet these needs.

Specific uses often vary between industries. Consider the following common scenarios:

- **Healthcare** - Research for treatment plans and preventative care is often based on data-based research. By de-identifying data, healthcare organizations can both meet HIPAA regulations and use the data for research purposes. Organizations can also use the data together with artificial intelligence to predict future health risks for current patients based on de-identified patient data, and then use the data to predict the most effective treatment options.
- **Education** - With 2020 being a census year, institutions of higher education are required to submit census data about students living on campus. However, FERPA requires institutions to protect student

data. De-identified data allows universities to meet both requirements without compromising the privacy of their students.

- **Banking** - With digital banking applications, customers and employees may encounter error messages – and developers must recreate and test the applications to solve the issues. However, because the data that's used qualifies as secure banking data, developers cannot use the actual data for testing purposes and stay in compliance with privacy laws. With de-identified data, financial institutions can legally use the data to recreate the errors.
- **Manufacturing** - Manufacturers ensure they are satisfying supply and demand by using year-over-year data to predict the correct number of specific products to produce. However, order information contains credit card data – which cannot legally be used. With de-identified data, businesses can more accurately predict future orders and comply with regulations.

Is the Resulting Data Unable to Be Tied to a Person?

One of the biggest concerns and challenges with de-identification is ensuring the resulting data is truly de-identified. Your company can be found in non-compliance if a person's identity can be reverse engineered, meaning someone can determine the person's identity from the de-identified data. Many businesses increasingly turn to technology solutions to manage the process and correctly strip the identifiers (both direct and indirect) from the data without removing the business value.

During the process, use the following 3-part test to achieve a "reasonable level of justifiable confidence," which is the bar set by privacy regulations:

1. **Take reasonable measures to ensure that data is de-identified**
2. **Publicly commits not to try to re-identify the data and**
3. **Contractually prohibits downstream recipients from trying to re-identify the data.**

After data is de-identified, you must ask the following questions to evaluate the effectiveness of the process:

1. **Is it still possible to single out an individual?**
2. **Is it still possible to link records relating to an individual?**
3. **Can information be inferred concerning an individual?**

Challenges Encountered During De-identification

While de-identification provides many benefits, businesses must also evaluate the challenges and proactively work to create the best process for de-identification.

- **Not adhering to data minimalism** - Because de-identification creates an additional data record, the process can add to data sprawl if not actively managed.
- **Ability to scale** - Data size can vary greatly, from gigabytes to petabytes – even data stored adjacent to each other. De-identification solutions must seamlessly manage disproportionate data sizes.
- **The time required to create data** - While it's possible to create a de-identified version of data using manual tools (such as Python), the process is time-consuming. Additionally, this method creates static data that does not update.
- **Knowing the schema updates** - De-identifying data such that it cannot be re-identified requires an understanding of all the elements in the data schema.

  Take the following real-world example:
  Two researchers at the University of Texas, Arvind Narayanan, and Professor Vitaly Shmatikov, were able to re-identity some portion of anonymized Netflix movie-ranking data with individual consumers on the streaming website. The data was released by Netflix 2006 after de-identification, which consisted of replacing individual names with random numbers and moving around personal details. The two researchers de-anonymized some of the data by comparing it with non-anonymous IMDb (Internet Movie Database) users' movie ratings. Very little information from the database, it was found, was needed to identify the subscriber. In the resulting research paper, there were startling revelations of how easy it is to re-identify Netflix users. For example, simply knowing data about only two movies a user has reviewed, including the precise rating and the date of rating give or take three days allows for 68% re-identification success.

  The overly specific data elements such as the exact user ratings were elements of the data schema that should have been removed. Knowing when a data set contains such elements is critical in the de-identification process.

- **Accurately synthesizing data** - One risk when de-identifying data is removing direct and indirect identifiers in a manner that retains the integrity and original meaning of the data. Businesses must ensure that their process does not add new issues into the data or skew the data, which provides false insights and answers when used for business purposes.

Turning to Technology for Automating the De-identification Process

When faced with the decision of remediation, data protection professionals should not automatically resort to solutions that appear to be the simplest and quickest – erasure through deletion, redaction, or shredding. By using de-identification as a remediation option, you can comply with all regulations while also retaining the utility of the data for future uses that help grow your company. Because manually de-identifying data is time-consuming, and it's easy to introduce errors, many companies automate the process through a data privacy technology solution that offers options for remediation – including de-identification.

To learn more about how Spirion can help you retain business value in your data, visit **https://www.spirion.com/products/data-privacy-manager/**

Talk to a Spirion data security and compliance expert today: expert@spirion.com

Spirion is the leader in data discovery, persistent classification, and protection of sensitive data on-premise and in the cloud. Since 2006, thousands of organizations worldwide have reduced their sensitive data footprint and proactively minimized the risks, costs and reputational damage of successful cyberattacks. Spirion provides greater command and control of sensitive data to leading firms across all industries from financial services to healthcare to public sector. Visit us at **spirion.com**

20200916 WP-DE-ID-001